10-28-2017

Statistical and Technical Methodologies for Duplicated Multiple-Samples Preference and Attribute Intensity Sensory Ranking Test

Kennet Mariano Carabante Ordonez
Louisiana State University and Agricultural and Mechanical College, kenneth.carabante@gmail.com

STATISTICAL AND TECHNICAL METHODOLOGIES FOR DUPLICATED MULTIPLE-SAMPLES PREFERENCE AND ATTRIBUTE INTENSITY SENSORY RANKING TEST

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The School of Nutrition and Food Sciences

by
Kennet Mariano Carabante Ordonez
B.S., Escuela Agrícola Panamericana. El Zamorano, 2008
M.S., Louisiana State University, 2013
December 2017

ACKNOWLEDGEMENTS

My deepest gratitude goes to my major advisor, Dr. Witoon Prinyawiwatkul, who has coached, guided and motivated me to be the best that I can be through this journey. His contributions to my work and development are unmeasurable. I am thankful to the Louisiana State University Agricultural Center, especially to Dr. William Richardson, Vice President of Agriculture at LSU for the financial support of my education. Because of this I feel committed to give back to the school in the future. I am also thankful to the members of my advisory committee, Dr. Marlene Janes, Dr. Bin Li, and Dr. Christopher Clark for their time and support throughout my PhD study.

I thank the entire food science faculty at the School of Nutrition and Food Sciences, from whom I learned at different stages of my career at LSU. My thanks go to Drs. Joan King, Jack Losso, Charles Boeneke, Evelyn Watts, Louise Wicker, Zhimin Xu, Subramanian Sathivel, Achyut Adhikari, and Kayanush Aryana. I am thankful to the LSU food incubator staff members Dr. Gabriela Crespo, Ashley Gutierrez and especially to my dear friend and brother, Dr. Marvin Moncada.

I could have not accomplished my goals at LSU without the support of my lab-mates and friends, Dr. Damir Torrico, Dr. Karen Garcia, Dr. Wannita Jirangrat, Kairy Pujols, Dr. Wisdom Wardy, Jose Alonso, Ryan Ardoin, Ana Ocampo, Amber Jack, Valentina Rosasco, Pitchayapat Chompracha, Yupong Gao, Jinjuta Jirawatjunya, as well as all the visiting scholars who supported my research. I also have a great deal of gratitude for all the members of Zamorano Agricultural Society, each one of them has helped me grow in one way or another. Especial thanks go to Juan Steer, Luis Vargas, Jorge Diaz, Alejandro Castro, Dr. Kevin Mis, Favio Herrera, and Dr. Daniel

Estrada, who were not only my room-mates but are dear friends. I also feel great gratitude towards Reynaldo Moreno and Ronald Maldonado for their continuous friendship and support.

This dissertation is dedicated to the memory of my parents, Silvia Ordoñez de Carabantes y Mariano Carabantes, you were and continue to be the main source of inspiration for my career and life. I also dedicate this research to all the members of the Carabante and Ordoñez families who have continuously loved me and supported me. I feel proud and honored to say I have three last names because Julio Sr., Guadalupe, Julio Jr., David, and Nancy Mazariegos have made me a part of their wonderful family. Above all I thank God for putting all these people in my way, allowing me to finally attain my Ph.D. degree.

.

TABLE OF CONTENTS

ABSTRACT

Ranking tests are important preference and attribute difference tools for sensory evaluation. Replicated testing is used widely to reduce the number of panelists required in other sensory methods such as discrimination. The information regarding replications sensory ranking is limited. This research evaluated important statistical and technical aspects for the development of the foundation for duplicated sensory ranking tests. Three studies were accomplished: 1) A study of nonparametric analyses on real preference ranked data; 2) a sensitivity study of two samples serving protocols for duplicated visual ranking, and 3) protocols comparison in taste. In study 1, 125 panelists ranked in duplicates each of two sets of three orange juice samples. One set contained very different samples and the other similar samples. Five methods of data analysis were evaluated. With similar samples, analyzing duplicates separately yielded inconsistent conclusions across sample sizes. The Mack-Skillings test was more sensitive than the Friedman test and is more appropriate for analyzing duplicated rank data.

Study 2 compared the sensitivity of duplicated yellow color intensity ranking served either in one or two sessions. Panelists (n=75) ranked both similar and different orange juice sets. For each set, rank sum data were obtained from (1) intermediate ranks from jointly re-ranked scores of two separate duplicates for each panelist, (2) joint ranked data of all panelists from the two replications in one serving session, and (3) median rank data of each panelist from two replications. Rank data (3) were analyzed by the Friedman test, while those from 1 and 2 by the M-S test. The similar-samples set had higher variation and inconsistency with one serving session, producing higher P-values than two serving sessions. Both M-S ranking protocols were more sensitive to color differences than Friedman on the medians.

For study 3, an identical design was used to evaluate both serving protocols of duplicated sweetness ranking tests. Separate duplicates were more sensitive for color but not in sweetness, especially with confusable samples. This showed that the conducting duplicated ranking in a single session can be beneficial, but it should be tested for the products and attributes of interest before standardizing testing.

CHAPTER 1. INTRODUCTION

1.1 Introduction

In sensory evaluation, ranking procedures help researchers to obtain analytical and affective information from the perception of subjects toward foods, personal care goods, cosmetics and many other consumer goods (Kemp and others, 2009). In sensory evaluation of foods, ranking tests require that each individual from a defined group of panelists rank three or more products, according to personal preference or perceived intensity of an attribute (Meilgaard and others, 2016). Panelists may be allowed to assign ties to closely perceived samples; however, it represents a different methodology than simple ranking and it has its own statistical analysis (Meilgaard and others, 2016). Without the ability to assign the same score to more than one sample, panelists are "forced" to order all samples from first to last or vice-versa. Therefore, this variant is commonly referred to as a forced choice multiple ranking test. The applications for the ranking tests are wide, but mostly help complement other sensory methods such as hedonic rating and intensity scaling screening from a large pool of products or as a direct source of information from special populations because of its simplicity (Lawless and Heymann, 2010). Other methods require ranking as a part of the screening exercises for panelist selection (Stone and Others, 2012) or the use of ranking combined with other scaling techniques such as in rank rating methods (Kim and O'Mahony, 1998).

Given the ordinal and dependent (within subjects) nature of the dataset obtained from a panel, the statistical testing of forced choice multiple rankings is accomplished with the Friedman (1937) non-parametric test (Gaito, 1980; Joanes, 1985; Lawless and Heymann, 2010; Meilgaard, 2016). The test has a null hypothesis and applications equivalent to a two-way Analysis of

Variance (ANOVA) without requiring normally distributed data. Panelists are used as complete blocks in a randomized complete block design (Lawless and Heymann, 2010).

The Friedman statistic follows a chi-squared distribution, which is obtained from a permutation of all the possible theoretical arrangements of the rank scores in a panel, and the likelihood of the observed compound difference when compared against that theoretical universe of permutations (Conover, 1999; Hollander and others, 2013). A limitation of the Friedman test is the inability to allocate replications of the complete rankings from the same panelists. The Friedman test can be used only after obtaining the median of the replicates because it requires independence between blocks (Conover, 1971). One of the main emphases for validity of sensory results is using a large enough number of panelists (Meilgaard and others, 2016). However, replications from the same panelists help account for intrapanelist variation due to the possible random assignation of scores in the absence of difference, also helping reduce the number of panelists and resources (Stone and others, 2012; Lawless and Heymann, 2010). Special statistical models were adapted for analysis of replicated preference and discrimination methods to determine if differences exists between two products (Ennis and Bi, 1999; Brockhoff, 2003). The Mack-Skillings test (1980) is extension of the Friedman procedure, capable of handling multiple replications of complete rankings from a block (Hollander and others, 2013). Replicated results equal those from the application of the Friedman's test, representing a viable option among other nonparametric tests for analysis of replicated sensory ranked data, e.g., the Van Elteren (1959) procedure.

1.2 Research justification

Replications are seldom used in raking tests and when used, the analysis with the Friedman test can be risky or inefficient. For discrimination, descriptive and simple preference tests there is

a solid literature foundation for replicated testing (Bi, 2006; Lawless and Heymann, 2010, Stone and Others, 2012; Meilgaard and others; 2016). Conversely, there is a clear gap in knowledge about handling replications in ranking tests. The availability of the M-S test can help the implementation of replicated ranking; however the methodology is to our knowledge, seldom known to sensory evaluation and consumer science. The adaptation of a replicated ranking methodology by researchers requires reliable answers to statistical and technical equally important concerns including: 1) Applicability, reliability, estimated power, benefits and possible compromises of the Mack-Skillings tests and competitor tests for statistical analysis; 2) Practical and measurable knowledge of the worthiness of applying replicated ranking tests; 3) Assessment of the impact of estimating *P* values for the M-S statistic for hypothesis rejection either with a chi-squared approximation or computer intensive methods; 4) How samples should be presented to panelists in a lab testing and if a break is required between replications as it pertains to sensory fatigue and other physiological effects.

Sensory evaluation uses humans as active instruments of measurement giving particular requirements in the design of experiments (Tormod and others, 2011). The comparison of methods or protocols for sensory testing (in this case replicated sensory ranking) usually requires applications on large enough panels to estimate power or sensitivity to differences (Kunert and Meyners, 1999; Garcia and others, 2012). Sensitivity to differences is one of the most desired qualities of sensory tests (Bi and Ennis, 1999). Sensitivity is affected by number of samples, training, instructions, categorical (or ordinal) decision strategy, order of presentation and statistical analysis among others (Bi, 2006). Research has covered several of these variables for multiple sensory tests. However, for duplicated ranking implementation, the consequences of the statistical analyses and if replicates could be served in the same joint ranking to a panelists are two variables

not previously studied. Only duplications are considered in this study because of possible sensory fatigue (Meilgaard, 2016). The effectiveness of a joint serving session for duplicated ranking might not be transferable from one sensory attribute to another, and in principle it can be harder to generalize the effectiveness of joint duplicated ranking to attributes perceived with different senses. Therefore, both serving protocols should be evaluated for different senses such as color vs. taste.

1.3 Research Objectives

This research aims to investigate aspects that consolidate the foundation of duplicated sensory ranking methodologies from statistical analysis to applications in preference and intensity ranking and possible serving protocols applied to tasks with different degree of difficulty. Namely, the objectives of this research are: 1) Evaluate the Mack-Skillings test and other alternative methods for statistical analysis of duplicated multiple samples preference ranking test; 2) Study the sensitivity to differences between the two possible serving protocols for multiple samples visual intensity ranking; 3) Evaluate the serving protocols for attribute intensity in a chemical sense, e.g., taste.

1.4 References

Bi J, Ennis, DM. 1999. Beta-binomial tables for replicated difference and preference tests. Journal of Sensory Studies, 14, 347-368.

Bi J, Ennis DM. 1999. The power of sensory discrimination methods used in replicated difference and preference tests. Journal of sensory studies. 14(3):289-302.

Bi J. 2006. Sensory Discrimination Tests and Measurements: Statistical Principles, Procedures and Tables. Blackwell Publishing; Ames, Iowa.

Brockhoff PB. 2003. The statistical power of replications in difference tests. Food Quality and Preference, 14, 405-417.

Conover W. 1971. Practical nonparametric statistics. John Wiley & Sons, Inc., New York.

Conover W. 1999. Practical nonparametric statistics. John Wiley & Sons, Inc., New York.

Friedman M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32(200): 675-701.

Gaito J. 1980. Measurement scales and statistics: Resurgence of an old misconception. Psychological Bulletin 87(3): 564-567.

Garcia K, Ennis JM, Prinyawiwatkul W. 2012. A large-scale experimental comparison of the tetrad and triangle tests in children. Journal of sensory studies 27(4):217-22.

Hollander M, Wolfe DA, Chicken E. 2013. Nonparametric Statistical Methods: New Jersey: John Wiley & Sons.

Joanes D. 1985. On a rank sum test due to Kramer. Journal of food science 50(5):1442-1444.

Kemp SE, Hollowood T, Hort J. 2011. Sensory evaluation: a practical handbook. John Wiley & Sons; West Sussex, England.

Kim KO, O'Mahony MI. 1998 A new approach to category scales of intensity i: Traditional versus rank-rating. Journal of Sensory Studies 13(3):241-249.

Kunert J, Meyners M. 1999. On the triangle test with replications. Food Quality and preference 10(6):477-82.

Lawless HT, Heymann H. 2010. Sensory evaluation of food: principles and practices: Springer Science & Business Media.

Mack GA, & Skillings JH. 1980. A Friedman-type rank test for main effects in a two-factor ANOVA. Journal of the American Statistical Association, 75(372), 947-951.

Meilgaard MC, Carr BT, Civille GV. 2006. Sensory Evaluation Techniques. Florida: CRC press. 448 p.

Stone H, Bleibaum R, Thomas HA. 2012. Sensory evaluation practices: Academic press.

Tormod N, Brockhoff PB, Tomic O. 2011. Statistics for sensory and consumer science. John Wiley & Sons.

Van Elteren PH. 1959. On the combination of independent two sample tests of Wilcoxon. Bulletin of the International Statistical Institute 37:351-361.

CHAPTER 2. LITERATURE REVIEW

2.1 Overview of sensory ranking tests

2.1.1 Introduction

Ranking is one of the most commonly used types of ordinal scale. The most direct approach is to ask subjects to arrange a set of products such that each succeeding product has more or less of intensity of an attribute or preference. With simultaneous product presentation, ranking is considered a direct method, and the products serve as their own frame of reference. The paired comparison (e.g., which sample is sweeter) and paired preference (which sample you prefer more) tests are a simplified case of the rank-order test and are of directional discrimination. When a large number of samples and time constraints are involved, it is not practical to use paired comparison tests. The multiple samples ranking test becomes useful for screening/presorting a large array of products to a smaller more manageable product subset. Data obtained from a multiple-samples ranking test are typically analyzed by the non-parametric Friedman's test. In some cases, in order to reduce the number of subjects, time and cost, duplicated ranking tests are performed, and data are analyzed using the non-parametric Friedman's test, not taking into consideration additional dependency between duplicates. Duplicated ranking testing can be beneficial provided that data analysis is properly handled (Carabante and others, 2016); however, this topic has not received much attention until recently.

This review discusses historical development of method and statistical analysis of sensory ranking tests, current practices and alternative procedures including duplicated ranking testing, some factors that induce errors, and statistical considerations for the duplicated multiple samples ranking test.

2.1.2 Simple paired preference test

According to Lawless and Heymann (2010), preference tests determine choices between two or more products by a group of panelists. The simplest preference comparison, based on two products, is known as the paired preference test. Each panelist simultaneously receives two samples (A and B) and is asked to identify which sample is more preferred. Because panelists must select one sample, it is a forced choice method. The two possible balanced serving sequences (AB or BA) should be randomized across a set of panelists. Advantages of this method include simplicity for consumers and simulation of actual consumer choice mechanisms (Lawless and Heymann, 2010). The test is suitable for use with children (Schraidt, 1991; Kimmel and Guinard, 1994). Moreover, it has been shown that illiterate panelists did not experience problems when performing the paired preference method (Coetzee and Taylor, 1996).

The main disadvantages of this method are a lack of absolute magnitude of differences and the results that may not associate with sensory liking. For example, a product "A" might be chosen over product "B", but consumers might dislike both products. In addition, Lawless and Heymann (2010) recommended avoiding a preference question right after other types of sensory discrimination tests, possibly due to pre-conceived frame of mind for sensory differences. Another drawback is a lack of appropriate handling for preference responses from panelists producing incorrect responses in discrimination; however, this issue was recently discussed (Rousseau and Ennis, 2017).

Data obtained from the forced choice paired preference test can be analyzed by statistical analysis methods with either a chi-square distribution, a normal distribution or a binomial distribution with probability of success (p) = 0.5. Using a binomial distribution, the probability of obtaining "y" selections for a product over another from "N" evaluations is expressed as: $p_y =$

$1/2^N \frac{N!}{(N-y)!y!}$. Bi (2006) provided tables of critical values based on a two-tailed test, showing the minimum number of responses favoring one product. With a large sample (N > 100), the binomial Cumulative Distribution Function (CDF) closely approximates the CDF of a standard normal distribution (Lawless and Heymann, 2010).

2.1.3 Variations of simple paired preference test

The simple paired preference test has been altered over the years to improve sensitivity to differences and increase power. The preference test with no-preference option or non-forced choice includes a third possible selection stating "no preference" or "equally preferred". For certain legal claims, this variation might be required (ASTM, 2006). According to Dhar (1997), difficulties in deciding among products can delay purchase decision, whereas opting for no preference or no choice can facilitate the process. There are four alternatives for handling data from non-forced preference tests: (1) a signal detection theory approach based on a Tau criterion and *d*' used in difference tests with no difference options (Braun and others, 2004); (2) a confidence interval approach used for large sample sizes (N > 100) and less than 20% of non-preference selections (Lawless and Heymann, 2010); (3) assigning of the non-preference selections equally to both products or based on the ratio of preference selections (Odesky, 1967); (4) elimination of the non-preference selections but using information about the description of the frequencies for the three options. Three common analytic methods (dropping, equal splitting, and proportional splitting) for handling no preference votes are compared with respect to power and type I error (Ennis and Ennis, 2012). They suggested that proportional splitting yielded more false alarms than expected and hence should not be used. Recently, a lack of appropriate data handling for preference responses from panelists failing to produce correct responses in discrimination tests was discussed (Rousseau and Ennis, 2017).

2.1.4 Multiple samples ranking tests

Ranking tests require panelists to completely rank a set of three or more samples for either general preference or the intensity of a specific attribute (Meilgaard and others, 2006). Simple ranking tests are regarded as a high-performance option for sensory analysis with the elderly (Wichchukit and O'Mahony, 2015). Among multiple sample tests, ranking tests are the cheapest, simplest and most efficient to set up, administer and perform (Stone and others, 2012). Nevertheless, carryover could generate interaction, and memory effects could become a confounding variable. Meilgaard and others (2006) recommended ranking tests for multi-sample evaluations with seven or less samples. Other than product testing, ranking tests have been used for panel performance or proficiency testing, as in the study by McEwan and others (2003), which required panelists to rank five apple juice samples according to their perceived sweetness. The applications for attribute intensity or difference ranking are wide. Some recent examples include difference tests for three tomato base samples (Belingheri and others 2015); consistency ranking of five samples of sweet potato porridge (De Carvalho and others, 2014); ranking of bitterness in three samples of spray-dried hydrolyzed casein (Subtil and others, 2014); and ranking of the taste and aroma attributes (terms) associated with the dissolved solids of fresh and dried lulo (*Solanum quitoense* Lam.) fruit samples (Forero and others, 2015); ranking of bitterness and pungency of six virgin olive oil samples to validate bitterness results from phenolic contents and bitterness index results (Aguilera and others 2015). Recent applications of preference ranking include a study by Karnopp and others (2015) on cookies containing whole-wheat flour and Bordeaux grape (*Vitis labrusca L.)* pomace. For all the previous examples, the statistical analysis performed was the non-parametric (distribution-free) test by Friedman (1937).

2.1.5 Variations of multiple samples ranking tests

The paired preference test is the two-sample version of a multiple-samples preference ranking test (Stone and others, 2012). More than two samples can be evaluated with the paired tests by grouping samples in pairs. Thus, the sensory evaluation or the analysis can be performed for all possible pairs or selected pairs. The Friedman (1937) test is used if each panelist evaluates all possible pairs. If only selected pairs are evaluated and different subjects were used for different pairs, a confidence interval tests is recommended (Bi, 2006). The "Q" statistic by Cochran (1950) serves as an alternative test for preference frequencies when the responses are dependent or matched, that is, all the subjects evaluate all the selected pairs.

Variations of multiple samples complete ranking tests can be applied to both preference and attribute intensity difference. The simplest ranking test does not allow ties; thus, panelists are "forced" to order all the samples. The Friedman (1937) test is the most widely accepted test for ranked data without ties from panelists in a Randomized Complete Block Design (RCBD). A ranking test variation allows panelists to assign ties between samples, thus affecting the statistical analysis. Hollander and Wolfe (1973) described an adjusted Friedman test for ranked data with ties from the RCBD setting.

A Balanced Incomplete Block Design (BIBD) is recommended for sensory or consumer studies with "too many" samples for a single subject to completely evaluate due to sensory fatigue, carryover or other physiological problems (Wakeling and McFie, 1995). The analysis of ranked data from a study carried out with a BIBD is performed with the test by Durbin (1951), which was later extended to more general incomplete block designs (Skillings and Mack, 1981). All these tests for the analysis of multiple samples ranked data are non-parametric, or distribution-free

methods that require fewer assumptions than tests based on standard-normal or parametric tests (Hollander and others, 2013).

2.1.6 Non-parametric or distribution-free tests

Statistical tests for the analysis of ranked data are usually non-parametric. Therefore understanding this class of statistics helps clarify why parametric ANOVA is not preferred. Hollander and others (2013) defined non-parametric methods as "statistical procedures that have certain desirable properties that hold under relatively mild assumptions regarding the underlying populations from which the data are obtained." In a simple metaphor, Conover (1999) described non-parametric statistics as "approximate solutions for exact problems". On the other hand, parametric statistics analogized "exact solutions to approximate problems". Non-parametric statistics differ from parametric even at the level of descriptive statistics. Boddy and Smith (2009) stated that when data are not normally distributed, the sample mean and standard deviation are not appropriate descriptive statistics of a population with a differently shaped distribution. A nonparametric alternative to the mean, i.e., the median, describes the center of a population. Because equal number of values lay below and above the median, the shape of the distribution loses importance.

Records of non-parametric statistics applications go back to the early 18^{th} century with the use of a sign test. However, mathematical approaches to assess the occurrence of an event, which were the foundation for the initial non-parametric tests date back to the renaissance (Bradley, 1968). Savage (1953) pointed to the year 1936 as the formal border between the use of certain tests of nonparametric resemblance and an understanding among statisticians that tests independent from the shape of a distribution should be available. One of the most important tests published after 1936 is the 2-way distribution-free ANOVA (Friedman, 1937). Since 1936 many parametric

alternative tests have been developed covering alternatives for one sample t test, two sample t test, one-way ANOVA, two-way ANOVA, correlation and regression, among many others (Hollander and others 2013).

2.1.7 Use of non-parametric statistics vs ANOVA

Comparing both classes of statistics can be difficult. Non-parametric advocates point out the advantages of non-parametric statistics; giving little credit to the robustness of parametric methods to deviations from normality. The advantages of non-parametric methods listed by Conover (1999) include: 1) less complex models; 2) fast and easy computation; 3) given that the development of non-parametric methods rarely used complex mathematics beyond algebra; someone able to understand the method is less likely to apply it when it is not required; 4) because of better use of information, non-parametric are more powerful than parametric statistics if the assumptions (or preconditions) of the latter class are not met. Nowadays with the use of statistical software, the second advantage becomes less important.

In addition, Hollander and others (2013) stated that the fast-paced advancement of nonparametric methods is also rooted in the following characteristics: 1) the ability to produce exact *P* values in tests, exact confidence intervals or confidence bands and exact error rates for multiple comparison procedures; 2) parametric methods are only slightly more powerful than non-parametric methods in conditions of normality; 3) resistance of outliers; 4) nonparametric methods can fit more data scales e.g., ranked data might not require original continuous data, such as in a ranking test of sensory evaluation; 5) availability of Bayesian non-parametric methods (Ferguson, 1973).

The study of handling non-normality is not exempt from contradictions; often related to the “robustness” of the parametric tests to deviations from normality. Bradley (1968) stated that

any deviation from normality produces a "non-exact result". The impact of the inexactitude will depend not only on the degree of non-normality, but also on other aspects including: area of rejection, shape of the sample distribution, variance size, variance homogeneity, an alpha level, sample size, relative characteristics of other samples etc. Originally, according to Bradley (1968), when sampling data were analyzed, contradictions occurred, for example: in several specific cases, smaller sample sizes showed less deviation from normality; less homogeneous variances yielded higher "normality". At that time, a sample size of $n > 4096$ was suggested to assure that deviations from normality delivered close to exact results, clearly not the current standards. Later, Bradley (1978) addressed that other authors (Boneau, 1960; Scheffé, 1959) not only failed to provide a numeric measure of "robustness", but promoted the term as an excuse for ignoring non-normality.

More recently, the robustness of parametric tests to handle deviations from normality received higher support. The approval was generally achieved with at least 10,000 simulated runs and for specific research fields; for example, in psychology (Rasch and Guiard, 2004); whereas, other studies, discuss specific alternatives. The Kruskal and Wallis (1952) and ANOVA tests represent one-way multiple-sample competitors. Khan and Rayner (2003) recommended ANOVA for small sample sizes ($n \leq 5$) even in non-normal conditions, whereas the Kruskal-Wallis outperformed ANOVA at large sample sizes and high Kurtosis. Additionally, Lantz (2013) recommended Kruskal-Wallis over ANOVA analysis when analyzing non-normal samples. Other options such as rank transformations and analysis under a parametric F distribution were also recommended by Conover and Iman (1981); however, specific restrictions apply regarding the distributional characteristics required. The selection of the appropriate class of statistics depends on many factors, including degree of non-normality, sample size, distributional shape, and kurtosis, number of treatments or tails. The literature is diverse and to avoid mistakes without

overcomplicating a choice it is important to research statistical method applications in the area of interest. Aside from non-parametric methods, generalized linear mixed models can help treat continuous, non- normal samples, adapting several distributions, but with ordinal data from low number of samples (products), non-parametric rank based tests are still the standard.

2.1.8 Tests of normality

Although the popularity of nonparametric tests has increased, Bradley (1968) claimed that the term "preconditions" fits better than "assumptions," which led to overuse of parametric statistics. For some researchers, "assumptions" implied that it should be assumed that in most cases, data are approximately normal or possess homogeneous variance, etc. Around that time, Bradley (1968) criticized the use of parametric statistics on sampled data that did not meet the "preconditions" of normal analysis. Shapiro and Wilk (1965) and Shapiro and others (1968) published a test for normality. The test is based on a correlation between the distribution of the data obtained and the scores of a normal distribution. It is considered the most powerful among normality tests (Steinskog and others, 2007; Ghasemi and Zahediasl, 2012). The proposed test uses the following null hypothesis (Ho): deviations from normality are not significant. If the test yields a rejection of the null hypothesis, Shapiro and Wilk (1965) suggested either to inspect the data of influential observations, data transformations or applying distribution-free methods.

Other than the test developed by Shapiro and Wilk (1965), Razali and Wah (2011) suggested that the tests by Kolmogorov and Smirnov (1933), Lilliefors (1967) or Anderson and Darling (1954) are also preferred over the sole use of graphical methods. Regarding power, these tests do not perform adequately for reduced sample sizes (30 or less), but for a larger sample size the Shapiro-Wilk's test is recommended (Razali and Wah, 2011; Yap and Sim, 2011). Over the years, generalizations of the normality tests for multivariate data also became available such as

those of Doornik and Hansen (2008) which can also perform with sample sizes as low as 10; Royston (1983); Villasenor-Alva and Estrada (2009), to name a few. Some attempts were done to improve the power of the Kolmogorov and Smirnov (1933) test by adjusting the proportions of the normal shape against which the data are compared (Drezner and others, 2010). The diversity of tests developed for normality evaluation indicates the growing emphasis of applying parametric tests only if the deviations from normality are not considered influential.

2.1.9 Friedman's test, the non-parametric RBD-ANOVA

Data from sensory multiple ranking tests rarely resemble normal distributions. Non-parametric techniques based on ranks serve to analyze original ordinal data sets and interval or continuous data with rank transformations (Kramer and others, 1974). The test by Friedman (1937) is the most widely recommended statistical analysis for ranked sensory data (Joanes, 1985; Chambers and Wolf, 1996; Meilgaard and others, 2006; Lawless and Heymann, 2010). The analysis tests the global null hypothesis (Ho: All $T_1 = T_2 = \dots T_k$, in preference or intensity) for more than k = 2 samples in a randomized block design (RBD) without block*sample interaction. Because the interaction effect is not tested (Hollander and others, 2013), factorial design effects are excluded. In RBD designs, panelists represent blocks; thus, requiring the two-way structure of the Friedman (1937) test. The analysis does not require previous interval data allowing the use of original ordinal ranked data from adults or children. Children can successfully perform preference (since age 3) and intensity (since age 4) rankings on multiple samples. On the other hand, intensity scaling is not recommended until age "6" (Guinard, 2000).

Rayner and Best (1990) recommended the test by Friedman (1937) over other 2-way nonparametric tests such as the Pearson (Cochran, 1952), Page (1963) and Anderson (1959) tests for taste testing data. The Friedman test is based on a two-way layout with model: X_{ij}: = $\mu + \beta_i + \tau_j + \varepsilon_{ij}$, without interaction where: μ is the overall mean (unknown); β_i is the effect of the i^{th} block

and τ_j is the effect of the j^{th} treatment or sample. The ε's are the mutually independent error variables originated from one continuous population (Hollander and Wolfe, 1973).

In multiple-sample ranking without ties, each panelist receives all "k" samples at once; assigning a unique rank (R [Xij]) value from 1 to k for each sample (Lawless and Heyman, 2010). The assigned values represent the order of attribute intensity or preference for the samples. The sum of the individual ranks from each sample, assigned by all the panelists (n) represents one of k-rank sums (Conover, 1971): $R_j = \sum_{i=1}^{n} R\left(X_{ij}\right)$. With degrees of freedom = k-1, the test for the Null Hypotheses Ho is: $Fr_{\chi^2}{}_{df=k-1,\alpha} = \left(\frac{12}{kn(k+1)}\right)\left[\sum_{j=1}^{k} R_j^2\right] - 3n(k+1)$

The null hypothesis is rejected if the statistic obtained is larger than the chi-squared critical value at α, and degrees of freedom k-1. The Friedman equation yields an asymptotically chi-squared statistic using the complete permutation structure of the rank scores assigned to a product by all panelists. For each data set, it subtracts the observed rank sums of each treatment to a mean rank sum followed by a sum of the squared differences (Joanes, 1985; Hollander and others, 2013).

2.1.10 Multiple comparison procedures for RBD designs

After rejecting the null hypothesis; multiple comparisons test paired differences between treatments. According to Lawless and Heymann (2010), either a non-parametric Tukey or LSD test are recommended. Additional critical value tables comparing corresponding values to both tests are discussed later. Hollander and Wolfe (1973) recommended the Tukey HSD analog, based on an experiment-wise error rate critical value rather than in the paired one used in the LSD. Conversely, Best (1990) claimed that the HSD method is highly conservative, proposing the use of the LSD non-parametric analog. However, he acknowledged that the HSD method avoids rejecting the null hypotheses with false differences. Both equations are shown below.

$\mathbf{LSD} = t_{\alpha/2\,\infty} * \sqrt{nk\,(k+1)/6} = LSD = z_{\alpha/2} * \sqrt{nk\,(k+1)/6}$ $\qquad$ $\mathbf{HSD} = q_{\alpha,k,\infty} * \sqrt{nk\,(k+1)/12}$

Where n = a number of panelists, k = a number of samples, $Z_{\alpha/2\,\infty}$ is a score from a standard normal distribution corresponding to one half of α for a two tailed comparison, $q_{\alpha,k}$ is the α^{th} distribution percentile for all "k" sample independent and normal variables

Other alternatives for multiple comparisons of the two-way layout exist. McDonald and Thompson (1967) provided tables of critical values; however, this method was not recommended for anything other than an experiment-wise error rate comparison (Church and Wike, 1979). In the same Monte Carlo study with k= 3, 5 or 7 and n = 8, 11 or 15, it is also recommended to avoid the test by Rhyne and Steel (1976) due to poor error rates performance. Among the other options, the Wilcoxon (1945) signed ranks test and a "stepped down" sign test obtained better error rates (Church and Wike, 1976).

2.1.11 Tables of critical values

In addition to tests that generate a statistic that leads to a *P* value, the analysis of sensory ranked data can be achieved through tables of critical values. These methods represent a quick alternative analysis to computing a non-parametric test. The tables show critical values for hypothesis rejection at a specific number of panelists "n" and a number of samples or treatments "k". The first set of tables was developed by Kramer (1956), based on the determination of all the possible rank sums, arranged in order from largest to smallest. All the rank sum values contained in the highest "1-α^{th}" percentile represented the rank sums that are significantly higher than the rest. All the rank sums contained in the lowest "α^{th}" percentile represented the rank sums that were significantly lower than the rest. The conservative nature of the tables, the lack of multiple comparisons inference, and the incorrect assumption of independence between the rank sums (Joanes 1985) motivated Newell and McFarlane, (1887) and Basker (1988) to create new tables. They simulated 10,000 panels for various n*k combinations, then obtained the highest rank sum difference form each panel to determine the largest absolute differences contained in a specific

"α^{th}" percentile on a contingency table. This method accounted for the dependence between samples given the inclusion of all the rank sum differences from al panelists. Nonetheless, Christensen and others (2006) declared such tables to be too conservative for multiple comparisons, but adequate for global hypothesis testing. They developed a new set of tables based not on the largest difference but on all the differences from each of the 10,000 simulated panels to construct the contingency tables that serve for obtaining the critical values for each "α", "n", and "k" values.

2.2 Replicated preference and difference tests

In sensory evaluation, the use of replicated preference and discrimination (difference) tests has mostly aimed to compare two original samples even if more samples are served to compare them, e.g., Triangle test. The study of replicated testing and analysis on multiple-samples tests such as ranking has received less attention. When properly analyzed, the use of replications in preference and discrimination testing is promoted to maximize the use of available panelists, reduce costs and improve statistical power (Lawless and Heymann, 2010). In addition, replication helps control intra-panelist variations, forcing panelists to re-assure decisions that could have arisen from randomness, and not from true perceptual difference (Stone and others, 2012). For such tests, the main concern has been the statistical analysis of data from the replications. According to Lawless and Heymann (2010), simple approaches include the analysis of replications separately, and based on diverse criteria, e.g., requiring both complete replications to be significantly different to declare a difference. Also, tabulate which panelists provided correct responses for all the replications performed and analyze the data based on a *Z* score test with an adjusted guessing probability for a specific test.

The need for extended information and less conservative analysis promoted analyses, which evaluated independence and/or over-dispersion of the set of data between the replicates, to assess if data from replicates can be pooled into one set. Smith (1981) described a method to test independence with a binomial test; in which if independence was achieved, i.e., overdispersion approaching zero, it would allow pooling the data form the replicates for analysis with a binomial test. This method could test independence but not the occurrence of patterns of agreement or disagreement between the replications within the data (Lawless & Heymann, 2010). If patterns exist they could inflate the variation for an originally binomial distribution expectation causing over-dispersion (Anderson, 1988). The beta-binomial model measures the occurrence of over-dispersion, and provides an adjustment for different levels, gaining increased popularity in discrimination testing (Harries and Smith, 1982; Ennis and Bi, 1998). The latest widely accepted adjustment to replicated discrimination testing is the corrected beta-binomial model (Brockhoff, 2003). Replicated testing is also recommended for descriptive tests (Stone and others, 2012), whereas duplications have also shown improvement in discrimination and reliability for product characterization with Check All That Apply (CATA) profiling and product spaces from projective mapping (Vidal and others, 2016). The last example used a long period (one week) between the duplicate assessments but they suggested that it could be done in a single session with a break after to minimize sensory fatigue.

2.2.1 Independence between and within panelists in ranking tests

Independence between blocks is a concern in both non-parametric and parametric statistical analysis (Mooijart and Bentler, 1991). In multiple-samples sensory ranking, each panelist should be independent and receive all “k” variables or samples to rank at once. Per Conover (1971), the blocks (b) in the Friedman (1937) test should be mutually independent; each composed of “k”

random variables representing the samples. Independence between blocks means that one block should not influence another block. When a panelist repeats a k-variate set of samples, and is accounted as another block, a high level of influence or dependence occurs. Stone and others (2012) stated that complete independence of judgments is utopic, but the risk of such dependence in parametric testing has not been clearly measured. The dependence between the judgements of a subject in a single ranking test without replications, and analyzed with the Friedman's test, is not undesirable and is accounted by a new assumption. Such assumption states that the scores for each sample evaluated should be equally likely under the null hypothesis, that is, when differences do not exist.

2.2.2 The Mack-Skillings test

Mack and Skillings (1980) proposed a non-parametric test alternative to the two-way ANOVA for one or more observations per block*sample combination (panelist*sample). The authors stated that the test is more powerful than an *F* test without a standard normal distribution, and almost as efficient under normality. The test is designed for an equal number of replications per cell or panelist*sample combination. Oron and Hoff (2006) affirmed that the Mack-Skillings (1980) test is a straightforward extension to the Friedman (1937), but it is much less known outside professionals of non-parametric statistics. With the Mack-Skillings test the new assumption persists, also requiring that all the scores regardless of replication should be equally likely (Hollander and others, 2013). The model of the test for a two-way with factors: α (rows or panelists) and θ (columns or samples) without interaction is:

$$Y_{ijk} = \mu + \alpha_i + \theta_j + E_{ijl}$$

i=1, j =1, and k = 1…$c_{ij} \geq 1$. Let: $N = \sum_{i=1}^{n} \sum_{j=1}^{k} cij = nck$, where: $\sum_{i=1}^{n} \alpha_i = \sum_{j=1}^{k} \theta_j = 0$, E_{ijk}'s are independent random variables, each with the same distribution function.

Based on that model, Hollander and others (2013) simplified the computation of Mack Skillings (M-S) statistic to: $M - S_{\chi^2}{}_{df=k-1,\alpha} = \left(\frac{12}{k(N+n)}\right)\left[\sum_{j=1}^{k} R_j^{*2}\right] - 3(N+n)$

Where, n= number of blocks (panelists in sensory evaluation), k = number of samples, c = number of complete replications for all n*k cells. $R_j^* = \sum_{i=1}^{n}\left[\sum_{l=1}^{c} {}^{r_{ijl}}/_{c}\right]$ = by-product rank sums (averaged from replications) of the within-block rankings which include all rank scores obtained from "nc" samples per panelist.

Table 2.1 Comparison of the Mack-Skilling and Friedman's test equations, parameters and multiple comparisons (MC)*

Characteristic	**Friedman**	**Mack- Skillings**
Number of Samples	k	k
Number of Panelists	n	n
Number of replications	Not available	c
Total observations	n*k	N= k*c*n
Samples ranked per panelist (vector size)	1 to k	1 to c*k
By sample rank sums	$R_j = \sum_{i=1}^{n} R\left(X_{ij}\right)$	$R_j^* = \sum_{i=1}^{n}\left[\sum_{l=1}^{c} {}^{r_{ijl}}/_{c}\right]$
Test equation	$\left(\frac{12}{kb(k+1)}\right)\left[\sum_{j=1}^{k} R_j^2\right] - 3b(k+1)$	$\left(\frac{12}{k(N+n)}\right)\left[\sum_{j=1}^{k} R_j^{*2}\right] - 3(N+n)$
Experiment-wise MC	$R_A - R_B \geq q_{\alpha,k} * \sqrt[2]{\frac{nk(k+1)}{12}}$	$R_A^* - R_B^* \geq q_{\alpha,k} * \sqrt[2]{\frac{k\,(N+n)}{12}}$
Paired-wise MC	$R_A - R_B \geq t_{\alpha,k,\infty} * \sqrt[2]{\frac{nk(k+1)}{6}}$	Not available

*j = the j[th] sample, i = the i[th] panelist and l = the l[th] replication.

Table 2.1 compares the parameters and characteristics of the M-S computation to those of the Friedman test. The M-S statistic asymptotically follows a Chi squared (χ^2) distribution with degrees of freedom (df) = k – 1. Nevertheless, a Monte Carlo simulation with 10,000 runs or an exact test was recommended by Hollander and others (2013) for less than 4 replications. With more replications, the Chi squared approximations yields slightly more conservative results. A

guide of R software codes for analysis of duplicated ranked data with a Monte Carlo simulation is available (Carabante and others, 2016).

The multiple comparison's procedure is based on an experiment-wise error rate, analog to a two-tailed HSD Tukey's test or Studentized range procedure for replicated data, with null hypothesis: H_0 = Sample A's rank sum (R_A) = Sample B's rank sum (R_B). Rejection of the Null hypothesis (H_0) is achieved when: $R_A^* - R_B^* \geq q_{\alpha,k} * \sqrt[2]{\frac{k\,(N+n)}{12}}$, where, $q_{\alpha,k}$ is the α^{th} distribution percentile for all "k" sample independent and normal variables (Mack and Skillings 1980). The Mack-Skillings test has also been evaluated on duplicated consumer preference ranked data, showing higher consistency than evaluating duplicates individually with the Friedman test and higher sensitivity than obtaining the medians of the replications (Carabante and others, 2016).

2.3 Factors affecting sequential sensory preference and difference tests

Given the active nature of real world perception and the variability of the human as an active instrument of measurement, biases or errors are unavoidable. Stone and others (2012) suggested that the straightforward approach to handle such factors and errors is to minimize them and balance their effect across all samples through awareness and design. The factors influencing sensory verdicts or judgement of panelists are mainly classified into: psychological and physiological. Very early physiological factors were considered errors (Guilford, 1954; Lawless and Heymann, 2010) and physiological factors are better defined by processes. The physiological processes affecting judgments included carryover (usually mitigated with randomized and balanced designing), sensory adaptation (O'Mahony, 1986), and memory (Amerine and others, 1965). In relationship with duplicated sensory ranking tests, this processes gains relevance if the duplicate sets of samples are served in the same joint ranking sessions. Whereas, with separate duplicates the "break" or inter duplicates time could also be affected.

2.3.1 Sensory adaptation

According to O'Mahony (1986), the human brain uses feature extraction and adaptation for protection from an overload of information. The first process involves removal of information, whereas adaptation attenuates the sensitivity of a sense to repetitive and redundant stimuli, also affecting subsequent stimuli over time (Wark and others, 2007). Sensory measurements are affected by adaptation when an input or stimulus remains constant, e.g., an odor or flavor. This sensation would generally vanish from the initial exposure and subsequent samples of the same general stimulus in multiple evaluation will be perceived as weaker in intensity (O'Mahony, 1986). This principle aids the notion that sensory evaluation of taste, smell and possibly vision, can benefit from a reduced number of evaluations by a panelist. In addition, adaptation requires less time to recover than fatigue since it is a sensory not a muscular process, gaining benefits from inter-trial breaks and rinsing to eliminate remaining stimulus. Nevertheless the occurrence of adaptation with a higher number of samples depends on the nature of the test and stimuli since initiation and duration are highly dependent on the stimulus (Köster, 2003), whereas other processes can also reduce sensitivity in analysis and interact with adaptation (O'Mahony, 1986).

2.3.2 Visual adaptation

The quickness of a ranking test can be beneficial for visual evaluations of a larger number of samples (Chambers and Wolf, 1996). Nevertheless, factors such as adaptation can impact a large sample set or a duplicated joint test. The most basic classification of adaptation mechanisms in visual perception describes mechanisms for chromatic, light and dark adaptation. According to Fairchild (2013), light adaptation is the decrease in sensitivity to changes in lightness due to high environmental illumination. For example, it is easier to see the stars at night than in the day when the sky illumination is several orders higher. Dark adaptation is the opposite response mechanism,

i.e., increasing visual sensitivity with higher environmental darkness, but it occurs slower than light adaptation (Kalloniatis and Luu, 2007).

Chromatic adaptation occurs with repeated exposure to a specific wavelength by the cones in the retina reducing sensitivity over time due to a lingering effect of the previous stimulus (Werner, 2014). It represents the changes in responsiveness of the three types of cone photoreceptors individually. The light and dark adaptation involves changes in all three types of receptors at once. Visual adaptation occurs through different mechanisms ranging from sensory exclusive, reflex-like or exclusively cognitive (Fairchild, 2013). Other forms of adaptation known as high level adaptation mechanisms are: spatial, frequency, contrast, motion adaptation, blur adaptation, noise adaptation, face adaptation and the McCollough effect (Clifford and Rhodes, 2005; Adams and others, 2010). Per Lawless and Heymann (2010), adaptation mechanisms must be considered when designing sensory tests and experiments, therefore visual adaptation must be considered for the design of replicated appearance and color evaluations of foods.

2.3.3 Memory implications in sensory testing

The impact of the memory of evaluators on the sensitivity to differences in sequential testing has been a subject of attention for both preference and discrimination tests. Ideal comparison in discrimination testing requires that the memory of the previous sample remains unaltered or undeteriorated when the subsequent samples are evaluated (Cubero and others, 1995). That is, when the panelist is still using immediate memory for the perception of the previous food, thus remarking the importance of inter-trial time reduction on memory decay. Nevertheless, it is important to consider that such inter-trial time reduction could be counterproductive preventing adaptation. Mantonakis and others, (2009) studied the sensitivity to differences in preference affected by the number of samples evaluated in the sequence (2 to 5). With a larger number of

samples, other factors rather than the wines themselves, showed higher effect on differences, e.g., position. They hypothesized that memory load and memory interference caused the reduction in preference sensitivity with more samples since, naive consumers tend to competitively analyze all samples to the previous favorite, resembling paired comparisons. With more samples tried, each new sample inserts interference through a new comparison.

When comparing the sensitivity of specific discrimination tests for testing perceived differences between two samples, Rousseau and others (1998) found that triadic tests or tests requiring the evaluation of three samples from two original treatment levels were less sensitive than a same different test which only requires two evaluations. The authors adjudicated the decrease in differentiation performance on memory decay given the longer time required for triadic tests with one more evaluation. Lau and others (2004) studied the specific impacts of memory decay (increased with longer inter-trial time) and memory interference (induced with the addition of additional samples or stimuli). Their results showed that memory interference was the more detrimental factor, but both can play roles in sensitivity reduction. Additional research on forced choice discrimination tests suggests that three sample tests (3AFC) were less sensitive than (2AFC) tests partly because of higher memory requirements (Dessirier and others, 1998; Roseau and O'Mahony (1997). In summary, the compendium of research suggests that memory is an important factor affecting sensitivity in difference or preference tests that require a larger number of samples and inter-trial rising.

2.4 Limitations of the ranking procedure

Some limitations of ranking tests include (Stone and Sidel, 1993):

- Typically, all products in a set of products must be evaluated before a judgment is made. This maximizes the potential for sensory fatigue and increases the likelihood of a loss in

differentiation among products. This problem is obvious when dealing with a large number of products or products with a lingering flavor/odor or greasiness or products with relatively small differences. Although ranking tests have wide applications, but with sample sets above three, they do not discriminate as well as tests based on the use of scales (Meilgaard and others, 2016).

- Because the ranking tests are directional, it is necessary to specify the characteristics and direction for the ranking. A problem occurs with untrained subjects, because they may not understand the specific characteristics (e.g., flavor intensity of earthy, muddy, musty from off-flavor catfish).
- Data provide no indication of the overall location of products on the attribute rated and no measure of the magnitude of differences between products.

2.5 Conclusion

The scientific discipline of sensory and consumer studies has expanded rapidly and now is equipped with new testing from improvements in discrimination methods, temporal perception, rapid descriptive methods, equivalence testing, measurement of emotions and wellness, impact of concepts, statements and sensory cues, applications on foods from insects, face recognition, eye tracking, noninvasive physiological methods among many others. During such evolution, the gap of duplicated ranking testing was not filled. Thus, postponing a possible improvement to one of the most straightforward methods of consumer presence and difference evaluation. The Mack-Skillings test suits the dependency between the samples and duplicates in a duplicated ranking, solidifying the foundation for testing. New studies (Carabante and others, 2016) suggest duplicating ranking tests in preference can potentially improve the consistency of the information and reduce the number of judges required.

2.6 References

Adams WJ, Gray KL, Garner M, Graf EW. 2010. High-level face adaptation without awareness. Psychological Science 21(2):205-210.

Aguilera MP, Jimenez A, Sanchez-Villasclaras S, Uceda M, Beltran G. 2015. Modulation of bitterness and pungency in virgin olive oil from unripe "Picual" fruits. European Journal of Lipid Science and Technology 117(9):1463-1472.

Amerine MA, Pangborn RM, Roessler EB. 1965. Principles of sensory evaluation of foods. Chap. 5, p. 249-254 1st ed. Academic press, London, New York.

Anderson DA. 1988. Some models for overdispersed binomial data. Australian Journal of Statistics 30(2):125-148.

Anderson R. 1959. Use of contingency tables in the analysis of consumer preference studies. Biometrics 15(4):582-590.

Anderson TW, Darling DA. 1954. A test of goodness of fit. Journal of the American Statistical Association 49(268):765-769.

ASTM. 2006. ASTM E1958-16A: Standard guide for sensory claim substantiation. ASTM international. Conshohocken, PA.

Basker D. 1988. Critical-values of differences among rank sums for multiple comparisons. Food Technology 42(2): 79-84.

Belingheri C, Ferrillo A, Vittadini E. 2015. Porous starch for flavor delivery in a tomato-based food application. LWT-Food Science and Technology, 60(1), 593-597.

Best D. 1990. Multiple comparisons for ranked data. Journal of Food Science 55(4):1168-1169.

Bi J. 2006. Sensory discrimination tests and measurements: Statistical principles, procedures and tables. Blackwell Publishing, Ames, IA.

Bi J, Ennis, DM. 1999. Beta-binomial tables for replicated difference and preference tests. Journal of Sensory Studies 14(3):347-368.

Boddy R, Smith G. 2009. Statistical Methods in Practice: For Scientist and Technologists. John Wiley & Sons, Ltd, Chichester, West Sussex, England.

Boneau CA. 1960. The effects of violations of assumptions underlying the t test. Psychological bulletin 57(1):49-64.

Bradley JV. 1968. Distribution-free statistical tests. 2nd ed. Prentice Hall, New Jersey.

Bradley JV. 1978. Robustness?. British Journal of Mathematical and Statistical Psychology 31(2):144-152.

Braun V, Rogeaux M, Schneid N, O'Mahony M, Rousseau BT. 2004. Corroborating the 2-AFC and 2-AC Thurstonian models using both a model system and sparkling water. Food Quality and Preference 15(6):501-507.

Brockhoff PB. 2003. The statistical power of replications in difference tests. Food Quality and Preference 14(15):405-417.

Carabante KM, Alonso-Marenco JR, Chokumnoyporn N, Sriwattana S, Prinyawiwatkul W. 2016. Analysis of Duplicated Multiple-Samples Rank Data Using the Mack–Skillings Test. Journal of food science 81(7):S1791-S1799.

Chambers IV E, Wolf MB. 1996. Sensory Testing Methods. ASTM, West Conshohocken, PA.

Christensen ZT, Ogden LV, Dunn ML, Eggett DL. 2006. Multiple comparison procedures for analysis of ranked data. Journal of food science 71(2):S132-S143.

Church JD, Wike EL. 1979. A Monte Carlo study of nonparametric multiple-comparison tests for a two-way layout. Bulletin of the Psychonomic Society 14(2):95-98.

Clifford CW, Rhodes G. 2005. Fitting the mind to the world: adaptation and after-effects in high-level vision, Oxford University Press.

Cochran WG. 1950. The comparison of percentages in matched samples. Biometrika 37(3/4):256-266.

Cochran WG. 1952. The $\chi 2$ test of goodness of fit. The Annals of Mathematical Statistics 23(3):315-345.

Coetzee H, Taylor J. 1996. The use and adaptation of the paired-comparison method in the sensory evaluation of hamburger-type patties by illiterate/semi-literate consumers. Food Quality and Preference 7(2):81-85.

Conover W. 1971. Practical nonparametric statistics. John Wiley & Sons, Inc., New York.

Conover W. 1999. Practical nonparametric statics. 2nd ed. John Wiley & Sons, Inc., New York.

Conover WJ, Iman RL. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. The American Statistician 35(3):124-129.

Cubero E, Avancini de Almeida TC, O'Mahony M. 1995. Cognitive aspects of difference testing: Memory and interstimulus delay. Journal of Sensory Studies 10(3):307-324.

De Carvalho IST, Tivana LD, Granfeldt Y. Dejmek P. 2014. Improved energy and sensory properties of instant porridge made from a roasted mixture of grated orange-fleshed sweet potatoes and flour made from shredded sun dried cassava. Food and Nutrition Sciences 5(14):1430-1439.

Dessirier J, Sieffermann J, O'Mahony M. 1999. Taste discrimination by the 3-afc method: testing sensitivity predictions regarding particular tasting sequences based on the sequential sensitivity analysis model. Journal of sensory studies 14(3):271-287.

Dhar R. 1997. Consumer preference for a no-choice option. Journal of Consumer Research 24(2):215-231.

Doornik JA, Hansen H. 2008. An omnibus test for univariate and multivariate normality. Oxford Bulletin of Economics and Statistics 70(s1):927-939.

Drezner Z, Turel O, Zerom D. 2010. A modified Kolmogorov–Smirnov test for normality. Communications in Statistics—Simulation and Computation 39(4):693-704.

Durbin J. 1951. Incomplete blocks in ranking experiments. British Journal of Mathematical and Statistical Psychology 4(4):85-90.

Ennis DM, Bi J. 1998. The beta-binomial model: accounting for inter-trial variation in replicated difference and preference tests. Journal of Sensory Studies 13(14):389-412.

Ennis JM, Ennis DM. 2012. A comparison of three commonly used methods for treating no preference votes. Journal of Sensory Studies 27(2):123-129.

Fairchild MD. 2013. Color appearance models, 3rd ed. John Wiley & Sons, Ltd, Chichester, West Sussex, England.

Ferguson TS. 1973. A Bayesian analysis of some nonparametric problems. The annals of statistics 1(2):209-230.

Forero DP, Orrego CE, Peterson DG, Osorio C. 2015. Chemical and sensory comparison of fresh and dried lulo (Solanum quitoense Lam.) fruit aroma. Food chemistry 169:85-91.

Friedman M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32(200):675-701.

Ghasemi A, Zahediasl S. 2012. Normality tests for statistical analysis: a guide for non-statisticians. International Journal of Endocrinology and Metabolism 10(2):486-489.

Guilford JP. 1954. Psychometric methods. McGraw-Hill, New York.

Guinard JX. 2000. Sensory and consumer testing with children. Trends in Food Science & Technology 11(8):273-283.

Harries J, Smith GL. 1982. The two-factor triangle test. International Journal of Food Science & Technology 17(2):153-162.

Hollander M, Wolfe DA. 1973. Nonparametric statistical methods. John Wiley & Sons, New York.

Hollander M, Wolfe DA, Chicken E. 2013. Nonparametric statistical methods, 3rd ed. John Wiley & Sons, New York.

Joanes D. 1985. On a rank sum test due to Kramer. Journal of food science 50(5):1442-1444.

Kalloniatis M, Luu, C. 2007. Light and dark adaptation. Retrieved from http://webvision.med.utah.edu/book/part-viii-gabac-receptors/light-and-dark-adaptation/.

Karnopp AR, Figueroa AM, Los PR, Teles JC, Simões DRS, Barana AC, Kubiaki FT, Oliveira JGBD, Granato D. 2015. Effects of whole-wheat flour and Bordeaux grape pomace (Vitis labrusca L.) on the sensory, physicochemical and functional properties of cookies. Food Science and Technology (Campinas) 35(4):750-756.

Khan A, Rayner GD. 2003. Robustness to non-normality of common tests for the many-sample location problem. Journal of Applied Mathematics & Decision Sciences 7(4):187-206.

Kimmel SA, Guinard J. 1994. Sensory testing with young children. Food technology 11:92-99.

Kolmogorov AN. 1933. Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari 4:83–91.

Köster EP. 2003. The psychology of food choice: some often encountered fallacies. Food Quality and Preference 14(5):359-373.

Kramer A. 1956. A quick, rank test for significance of differences in multiple comparisons. Food Technology 10(8):391-392.

Kramer A, Kahan G, Cooper D, Papavasiliou A. 1974. A non-parametric ranking method for the statistical evaluation of sensory data. Chemical Senses 1(1):121-133.

Kruskal WH, & Wallis WA. 1952. Use of ranks in one-criterion variance analysis. Journal of the American statistical Association 47(260):583-621.

Lantz B. 2013. The impact of sample non-normality on ANOVA and alternative methods. British Journal of Mathematical and Statistical Psychology 66(2):224-244.

Lau, S., O'Mahony, M., & Rousseau, B. (2004). Are three-sample tasks less sensitive than two-sample tasks? Memory effects in the testing of taste discrimination. Perception & psychophysics 66(3):464-474.

Lawless HT, & Heymann H. 2010. Sensory evaluation of food: principles and practices. Springer Science & Business Media, New York.

Lilliefors HW. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association 62(318):399-402.

Mack GA, & Skillings JH. 1980. A Friedman-type rank test for main effects in a two-factor ANOVA. Journal of the American Statistical Association 75(3):947-951.

Mantonakis A, Rodero P, Lesschaeve I, & Hastie R. 2009. Order in choice: Effects of serial position on preferences. Psychological Science 20(11):1309-1312.

McDonald B, Thompson W. 1967. Rank sum multiple comparisons in one-and two-way classifications. Biometrika 54(3/4):487-497.

McEwan JA, Heiniö RL, Hunter EA, Lea P. 2003. Proficiency testing for sensory ranking panels: measuring panel performance. Food Quality and preference 14(3):247-256.

Meilgaard MC, Carr BT, Civille GV. 2006. Sensory Evaluation Techniques, 4th ed. CRC Press, Boca Raton, FL.

Meilgaard MC, Carr BT, Civille GV. 2016. Sensory Evaluation Techniques, 5th ed. CRC Press, Boca Raton, FL.

Mooijaart A, Bentler PM. 1991. Robustness of normal theory statistics in structural equation models. Statistica Neerlandica 45(2):159-171.

Newell G, & Macfarlane J. 1987. Expanded tables for multiple comparison procedures in the analysis of ranked data. Journal of Food Science 52(6):1721-1725.

O'Mahony M. 1986. Sensory adaptation. Journal of Sensory Studies 1(3-4):237-258.

Odesky SH. 1967. Handling the neutral vote in paired comparison product testing. Journal of Marketing Research 4(2):199-201.

Oron AP, Hoff PD. 2006. Kruskal-Wallis and Friedman type tests for nested effects in hierarchical designs. Centre for statistics and the social science, University of Washington. Retrieved from https://www.csss.washington.edu/Papers/2006/wp68.pdf.

Page EB. 1963. Ordered hypotheses for multiple treatments: a significance test for linear ranks. Journal of the American Statistical Association 58(301):216-230.

Rasch D, Guiard V. 2004. The robustness of parametric statistical methods. Psychology Science 46(2):175-208.

Rayner J, Best D. 1990. A comparison of some rank tests used in taste-testing. Journal of the Royal Society of New Zealand 20(3):269-272.

Razali NM, Wah YB. 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics 2(1):21-33.

Rhyne A, Steel R. 1967. A multiple comparisons sign test: all pairs of treatments. Biometrics 23(3):539-549.

Rousseau B, Ennis DM. 2017. Preference without a difference. IFPress 20(1):3-4.

Rousseau B, Meyer A. O'Mahony M. 1998. Power and sensitivity of the same-different test: comparison with triangle and duo-trio methods. Journal of Sensory Studies, 13(2), 149-173.

Rousseau B, & O'Mahony M. 1997. Sensory difference tests: Thurstonian and SSA predictions for vanilla flavored yogurts. Journal of Sensory Studies, 12(2), 127-146.

Royston J. 1983. Some techniques for assessing multivariate normality based on the Shapiro-Wilk. Applied Statistics 32(2):121-133.

Savage IR. 1953. Bibliography of nonparametric statistics and related topics. Journal of the American Statistical Association 48(264):844-906.

Scheffe H. 1959. The analysis of variance. John Wiley & Sons, New York.

Schraidt MF. 1991. Testing with children. ASTM Standardization News 19(3):4245-4248.

Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality (complete samples). Biometrika 52(3/4):591-611.

Shapiro SS, Wilk MB, Chen HJ. 1968. A comparative study of various tests for normality. Journal of the American Statistical Association 63(324):1343-1372.

Skillings JH, Mack GA. 1981. On the use of a Friedman-type statistic in balanced and unbalanced block designs. Technometrics 23(2):171-177.

Smith GL. 1981. Statistical properties of simple sensory difference tests: confidence limits and significance tests. Journal of the Science of Food and Agriculture 32(5):513-520.

Steinskog DJ, Tjøstheim DB, Kvamstø NG. 2007. A cautionary note on the use of the Kolmogorov–Smirnov test for normality. Monthly Weather Review 135(3):1151-1157.

Stone H, Sidel JL. 1993. Sensory evaluation practices, 2nd ed. Academic press, New York.

Stone H, Bleibaum R, Thomas HA. 2012. Sensory evaluation practices, 4th ed. Academic press, New York.

Subtil S, Rocha-Selmi G, Thomazini M, Trindade M, Netto F Favaro-Trindade C. 2014. Effect of spray drying on the sensory and physical properties of hydrolyzed casein using gum Arabic as the carrier. Journal of food science and technology 51(9):2014-2021.

Vidal L, Jaeger SR, Antúnez L, Giménez A, Ares G. 2016. Product spaces derived from projective mapping and CATA questions: Influence of replicated assessments and increased number of study participants. Journal of Sensory Studies 31(5):373-381.

Villasenor Alva JA, Estrada EG. 2009. A generalization of Shapiro–Wilk's test for multivariate normality. Communications in Statistics—Theory and Methods, 38(11), 1870-1883.

Wakeling IN, MacFie HJ. 1995. Designing consumer trials balanced for first and higher orders of carry-over effect when only a subset of k samples from t may be tested. Food Quality and Preference 6(4): 299-308.

Wark B, Lundstrom BN, Fairhall A. 2007. Sensory adaptation. Current Opinion in Neurobiology 17(4):423-429.

Werner A. 2014. Spatial and temporal aspects of chromatic adaptation and their functional significance for colour constancy. Vision research 104:80-89.

Wichchukit S, O'Mahony M. 2015. The 9-point hedonic scale and hedonic ranking in food science: some reappraisals and alternatives. Journal of the Science of Food and Agriculture 95(11):2167-2178.

Wilcoxon F. 1945. Individual comparisons by ranking methods. Biometrics bulletin 1(6):80-83.

Yap B. & Sim C. 2011. Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation, 81(12):2141-2155.

CHAPTER 3. ANALYSIS OF DUPLICATED MULTIPLE-SAMPLES RANK DATA USING THE MACK–SKILLINGS TEST

3.1 Introduction

A multiple-samples ranking test is a simple and essential tool for sensory discrimination in terms of preference and/or attribute intensity. It is simple, quick, and friendly to untrained consumers (Lawless and Heymann 2010). Rank data are inherently ordinal; hence they should be analyzed by nonparametric statistical analysis (Bi 2006). The Friedman rank sum test is perhaps the most commonly used method for analysis of rank preference data. Replicated preference test may increasingly gain relevance because it increases the number of replications per sample and hence reduces cost of sensory testing. When there is more than one replication within a block, and the number of replications is equal for all samples, the Mack–Skillings test can be used for the global null hypothesis testing of no differences among samples. The testing of the global null ranking hypothesis ($H0$: all samples are not different or $H0$: $t_1 = t_2 = \ldots = t_k$) normally takes two main routes: nonparametric (distribution-free) analysis of variance (ANOVA) or ready-to-use tables of critical values which provide hypothesis test conclusions but not a degree of significance via a P value. The tables of critical values for rank analysis were first developed by Kramer (1956). Based on those tables, other versions and extensions were developed by Bradley and Kramer (1957), Kramer (1960, 1963), and Kahan and others (1973). Kramer's method cannot provide multiple paired comparisons among samples, only determining if each individual sample is categorized into either "significantly lower," "significantly higher," or "not different from the rest." This limitation is explained by the nature of the tables which categorizes rank sums using the permutation distribution of all possible rank permutations $\{(k)^n\}$, where k is the number of treatments and n is the number of panelists, and determines a critical value cutoff at α.

This chapter originally appeared in Journal of Food Science Volume 81, issue 7, pages: S1791-S1799 in July 2017.

Reprinted by permission

To determine the "significantly lower" group, cut-off rank sums are selected after locating the highest absolute rank sum value that falls within 0 and α. The selected critical value for the "significantly higher" group is the lowest rank sum found between $(1 - \alpha)$ and 1. This construction incorrectly assumes that sample's rank sums are independent, which is a reason to motivate different alternatives (Joanes 1985). Newell and MacFarlane (1987) created, whereas Basker (1988) expanded critical value tables using the highest simulated (n = 10000 simulations) absolute rank sum difference for fixed sets of samples and block (panel) sizes at α = 0.1, 0.05, and 0.01. Later, Christensen and others (2006) suggested that the Basker's method originally created for multiple comparisons was better suited for global hypothesis testing given its conservative approach.

They created new table sets for multiple comparisons (that is, the LSD Test), using only simulated paired differences instead of the range. Among the distribution-free tests, the Friedman's test for several related samples is a 2-way ANOVA analog (Friedman 1937; Conover 1971; Hollander and Wolfe 1973), where, in sensory research, panelists represent complete blocks (RBD design without treatment × panelist interaction). The test is recommended for ranked preference analysis by Joanes (1985), Meilgaard and others (2006), and Lawless and Heymann (2010). A preference test with replications involves panelists participating more than once in the same study, evaluating the exact same set of samples. These tests require special statistical analyses that account for the nonindependence of the data. For laboratory or central location test (CLT), replicated preference tests are not common, but using replications correctly can reduce the cost of recruiting, screening, and transportation of panelists (Lawless and Heymann 2010). Consumer responses can change from one replication to another, and accounting for this intrapanelist variation is necessary. In a nonreplicated paired preference test, Cochrane and others (2005) stated

that it is difficult to determine if consumer response is based on true preference or the inherent randomness of the forced decision when a truly preferred product is not found by a panelist. Hence, they recommended replicated paired preference testing to be done.

In sensory evaluation, data from some replicated sensory tests including discriminative and paired preference tests are analyzed with β-binomial (Ennis and Bi 1998) or corrected β-binomial models (Brockhoff 2003). Both tests included the overdispersion between replications (Anderson 1988) to increase testing power (Bi 2006). When evaluating the suitability of a distribution-free method to fit the researcher's needs, several aspects are considered. Important evaluations using Bootstrap and Monte Carlo simulations of power and asymptotic relative efficiency (Pitman 1936) are without question valuable and help sensory scientists decide between statistical tests. Practical applications of statistical methods in real-life situations, including actual consumers from feasible panel sizes, can also help determine method selections. According to Brockhoff and Schlich (1998), researchers compensate the lack of panelists by having them replicate the discrimination test several times. However, suitable data analysis for the replicated multiple-samples rank data is less known and applied. The Mack–Skillings procedure (Mack and Skillings 1980) based on proportional frequencies represents an extension to the Friedman's distribution-free test to analyze more than 1 replication per treatment–panelist (block) combination. Each repetition of the complete ranking test by a panelist represents 1 additional data cell for each treatment-panelist combination. The method is also explained by Hollander and others (2013), and multiple comparison procedures are provided in both sources (Mack and Skillings 1980; Hollander and others 2013).

The Mack–Skillings (1980) test has not received sufficient consideration in sensory and consumer sciences. The description and application of this test can help researchers make more

informed decisions. Therefore, the objectives of this study were to explore the use of the Mack–Skillings test for analysis of duplicated multiple-samples preference rank data, and to compare the results with those analyzed by the Friedman's test. Furthermore, the analysis was done to demonstrate effects of degree of product divergence (different-sample vs. similar-sample sets) and sample size (n = 10 to 125). In addition, to explain the Mack–Skillings computation, a brief example was described in the section "Materials and Methods."

3.2 Materials and methods

3.2.1 Sample description

Two sets of 3 orange juice samples each were designed to produce a different-samples set (Set 1), which was expected to give higher absolute differences among the 3 samples than a similar-samples set (Set 2). Both sets included one sample of 100% orange juice without pulp (Tropicana Products, Inc., Chicago, Ill., U.S.A.). Set 1 was completed with 2 dilutions of 100% Tropicana orange juice with purified spring water (w/w) to obtain 70% and 40% orange juices. Similarly, Set 2 was completed with 2 samples containing 95% and 90% orange juices.

3.2.2 Multiple-samples ranking tests

The research protocol for this study was approved (IRB# HE 15 to 9) by the Louisiana State Univ. (LSU) Agricultural Center Institutional Review Board. A group of 125 panelists was recruited from a pool of faculty, staff, and students at the LSU campus. The criteria for recruitment were: availability and no allergy for orange juice. Those who self-indicated sensory deficits (ageusia and/or anosmia) were excluded from this study. They were asked to rank 3 samples without giving ties (1 = most preferred and 3 = least preferred). All panelists completed the duplicated preference ranking test of both sample sets (S1 on 1 day and S2 on another day). They took a 15-min mandatory break between the 2 replications. They were asked to step out of the

sensory partitioned booth room, wait in the reception area, and then repeat the test on a biological replication of the set (different aliquots of the same orange juice products). Different random 3-digit sample codes were used to avoid biases between the 2 replications and the 2 sample sets. For all 4 individual ranking tests (2 sample sets × 2 replications), samples were presented in a counter-balanced arrangement. Panelist identifications were recorded to ensure the matching of the replications data analysis. The test room was illuminated with cool, natural, and fluorescent lights. Crackers, water, and expectoration cups were provided to consumers to use to minimize any residual effects between samples. The Compusense® 5 release 5.6 (Compusense Inc., Guelph, ON, Canada) software was used to develop the questionnaire and collect the data.

3.2.3 Ranking statistical analysis alternatives

The planned data structure (k= 3 treatments, n =125 panelists and c= 2 replications) allowed several alternative analyses varying in data handling or the test used. All the analysis performed asymptotically followed a chi-square distribution with degrees of freedom (df) = k-1 = 2, enabling direct contrast or comparison of chi-square statistics. The four approaches of data analysis used in this study are described as follows:

1. Averaging the rank sums of both replications followed by the Friedman's test at several sample sizes (n = 10-125 panelists). Hollander and others (2013) pointed out that using the Friedman's (1937) test after obtaining the median of rank scores from the replications is a more conservative alternative non-parametric analysis of replicated rank data. In this study in which c = 2 replications, the averaged rank sums of the replications (of "n" panelists) by sample equals the sum of the median scores of the replications by sample.
2. Data analysis involved individual replications separately analyzed using the Friedman's test.
3. The Mack-Skillings procedure was applied on both replications jointly.

4. The Friedman's test was performed on data pooled from both replications, transforming "n" into "2n = n non-independent blocks" to emulate an analysis that violates the assumption of independence between blocks (ranks from the same panelist are used as individual blocks).

3.2.4 On the Mack-Skillings test

The distribution-free Mack-Skillings (1980) test is an asymptotically chi-squared test for general hypothesis testing of the RBD design with more than one observation per cell (block-treatment combination). In a traditional Friedman's test data arrangement, the treatments represent columns (*j*), and the panelists or blocks represent rows (*i*), restricting to one observation per each cell. In the Mack-Skillings test, each block contains all rank data from all replications; this test is exemplified by Hollander and others (2013). While its asymptotic relative efficiency was praised by Rinaman Jr (1983) in terms of power, a higher asymptotic relative efficiency means more power when cell size is fixed and the number of blocks become large or vice versa. The Mack-Skillings chi-square statistic is calculated as follows:

MS $=\left(\frac{12}{k(N+n)}\right)\sum_{j=1}^{k}\left(R_j\frac{N+n}{2}\right)^2 = \left(\frac{12}{k(N+n)}\right)\left[\sum_{j=1}^{k}R_j^2\right]-3(N+n)$, where "n" = the number of panelists, "k" = the number of treatments, "c" = the number of complete ranking replications, "N = nkc" and R_j = the by-product rank sums (averaged from replications) of the within-block rankings which include all rank scores obtained from "nc" samples per panelist.

An experiment-wise multiple comparisons procedure is also available for the Mack-Skillings test (Mack and Skillings 1980; Hollander and others 2013) with a null hypothesis: H_0 = Sample A's rank sum (R_A) = Sample B's rank sum (R_B). Reject H_0 if:

$$R_A - R_B \geq q_{\alpha,k} * \sqrt[2]{\frac{k\,(N+n)}{12}},$$

Where, $q_{\alpha,k}$ represents the α^{th} distribution percentile for all "k" sample independent and normal variables (Mack and Skillings 1980). This test relies on an experiment-wise error rate, i.e., the HSD analog procedure. Although this method was recommended by Hollander and others (2013), they also provided a conservative multiple comparisons test based on the Scheffé approximation. According to this procedure, two samples are significantly different if their absolute deference is greater than or equal to the critical value, as follows:

$$|R_A - R_B| \geq \sqrt{[k(N+n)ms_\alpha /6]}$$

Where, ms_α is the variable upper tail critical value given the number of panelists, replications and test samples. The R codes for calculating critical values and the Mack-Skillings statistic are available in Figure 3.1.

3.2.5 Assumption of independence between blocks and sensory fatigue

Although the research on "non-independence" in the Friedman's test is sparse, this specific assumption "independence between blocks" is of importance for accurate analysis (Rigdon 1999). Applying the Mack-Skillings test to analyze replicated preference rank data helps to avoid a violation to this assumption; however, a new point of consideration arises. Both Friedman's and Mack-Skillings tests replace their within-blocks independence assumption by an assumption (null hypothesis) that all $(k!)^n$ rank matrix configurations composed of all individual rank scores are equally likely for the Friedman's test, while all $[(ck)!]^n$ rank scores configurations are equally possible for the Mack-Skillings test (Hollander and others 2013).

According to the Mack-Skillings test (Hollander and others 2013), when k =3 and c =2, there are two practical serving protocols: each panelist ranks all kc = 6 samples in one session, resulting in a set of 6 rank scores (1, 2, 3, 4, 5, and 6), and/or each panelist ranks the same k=3 samples twice (i.e., in two separate sessions) and the combined data for each panelist consist of

three sets of ties (i.e., 1, 1, 2, 2, 3, 3), hence averaging within-block rank scores (intermediate ranks for two replicates) for each panelist is required prior to data analysis. Regardless of the serving protocols, the individual rank scores (1, 2, 3, 4, 5 and 6) or intermediate rank scores will ultimately contribute to an average by sample rank sum of "c" replications ($R_j^* = \sum_{i=1}^{n} r\left[\sum_{l=1}^{c} r_{ijl} / c\right]$) (Hollander and others 2013), where the averaged rank sum is derived from the sum (up to n) of each of the averaged rank scores. Such averaged rank scores are the sum of all scores from the i^{th} block for the sample *j*, divided by c).

In a typical multiple-samples preference test, panelists rank all samples at once and re-tasting is permitted (Stone and Sidel 1993). In a sensory research scenario where panelists serve as active instruments without sensory fatigue, two replications (c = 2) could be evaluated jointly in a single session. As such, each panelist would rank all "kc" samples served at once (the identity of the samples should not be revealed); therefore, under the null hypothesis, each r_{ij} rank score is equally likely for each *i* panelist. However, in a more practical and realistic situation, and from a sensory fatigue standpoint, ranking "k" samples twice with a resting period in between is more favorable. Extending c > 3 could also involve sensory fatigue so c should be kept minimal. Re-ranking two sets of rank scores from the same sample set and the same panelist could be thought of as a rank transformation to a single block. Rank transformation is normally employed when data intended for parametric ANOVA analysis do not meet the normality assumption and that deviation from normality could not be handled by ANOVA's robustness. Only in such a case, rank transformation offers more sensitivity to treatment effects than ANOVA (Edgington, 1980). In our current study, each panelist ranked the same k samples twice with a 15-min mandatory break in between, and both sets of rank scores were ordinal and non-normal, and intended to be re-ranked jointly within a block as in an RBD-rank transformation.

To demonstrate how the Mack-Skillings procedure works, an example is given below. When k = 3 and c =2, each panelist will provide a total of six joint rank scores. The following example explains the Mack-Skillings procedure when k = 3 (A, B and C), c = 2, and n = 5 panelists, including data structure (Table 1). The first half of Table 1 (left side) illustrates a matrix arrangement of the original data set from a duplicated ranking test.

Table 3.1 An example of the Mack-Skillings re-ranked data from n = 5 panelists, c = 2 replications and k = 3 treatments

Obtained data (k=3, n=5, c=2)							Averaged rank data to accommodate ties						
n	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**	**n**	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**
1	2	1	3	3	1	2	1	3.5	1.5	5.5	5.5	1.5	3.5
2	1	1	3	3	2	2	2	1.5	1.5	5.5	5.5	3.5	3.5
3	1	1	3	3	2	2	3	1.5	1.5	5.5	5.5	3.5	3.5
4	1	1	2	3	3	2	4	1.5	1.5	3.5	5.5	5.5	3.5
5	3	1	2	2	1	3	5	5.5	1.5	3.5	3.5	1.5	5.5

A, B, and C are treatments. 1 and 2 indicates replication.

The first step is to compute the averaged within-block rank scores (intermediate ranks) obtained from each of the "kc" presented samples to accommodate ties as seen on the second half of Table 1 (right side). Next compute Rj for all 3 products as follows:

For A, Rj = (3.5+1.5+1.5+1.5+5.5+1.5+1.5+1.5+1.5+1.5)/2 = 10.5

For B, Rj = (5.5+5.5+5.5+3.5+3.5+5.5+5.5+5.5+5.5+3.5)/2 = 24.5

For C, Rj = (1.5+3.5+3.5+5.5+1.5+3.5+3.5+3.5+3.5+5.5)/2 = 17.50

Then plugging in values in the Mack-Skillings chi-square equation as follows:

$$\left(\frac{12}{3(30+5)}\right)\,[[10.5]^2 + [24.5]^2 + [17.5]^2] - 3(30+5) = \ 11.2$$

With df = k-1 = 2, and a critical value of 5.99; *P* value = 0.0037. The null hypothesis (Ho: A=B=C) is rejected, and in conclusion, at least one pair of samples is different.

Furthermore, the multiple comparisons are calculated using: $|R_A - R_B| \ \geq q_{\alpha,k} * \sqrt[2]{\frac{k\,(N+n)}{12}}$

The critical value is obtained as follows: $= \quad q_{0.05,3} * \sqrt[2]{\frac{3\,(30+5)}{12}} \quad = 3.315 * 2.9584 =$ **9.8059**

Then, for each paired comparison :

$|R_A - R_C|$ = 10.5 - 17.5 = -7; |-7| = 7 < 9.8059) = Failure to reject the Null hypothesis (A = C).

$|R_B - R_C|$= 24.5 - 17.5 = 7 < 9.8059) = Failure to reject the Null hypothesis (B = C).

$|R_A - R_B|$= 10.5 - 24.5 = -14; |-14| = 14 ≥ 9.8059) = Reject the Null hypothesis (A ≠ B).

Therefore, we concluded that A and B were the only significantly different pair of samples.

3.3 Results

For both sample sets (S1 and S2), data were analyzed at varying sample sizes (n) from 10 to 125; the smaller n was created by random selection from n = 125 (Tables 2 and 3). At any given "n", the rank scores from the same randomly selected panelists were analyzed using the four data analysis methods mentioned earlier. At all "n" sizes, it was verified that the same panelists composed the blocks across replicates and data analysis methods.

3.3.1 Effect of sample size on chi-square and *P* values

With the different-samples set (Table 2), increasing "n" generally increased the chi-square values while decreasing the corresponding *P* values (except one case, where n = 30-35 for the individual replication 1). Without exception, the null hypothesis was rejected at all sample sizes and data analysis methods. In addition, the *P* values showed a high degree of significance across all analysis methods ($P < 0.0002$, except one case at n = 10 where $P = 0.0055$). The Mack-Skillings test was relatively more sensitive to the differences (higher chi-square and lower *P* values) at all "n" sizes. Overall, for samples that were very different (less variation in rank data from the two replications from each panelist), sample size and data analysis methods may be less critical as they provided consistent results of the main effects, when compared with the similar-samples set (Table 3) as demonstrated in this study. With the similar-samples set (Table 3), more variations in the

obtained rank data from the two replications were observed. An increase in "n" did not always yield higher chi-square values and lower corresponding *P* values, especially with the individual replications. For instance, for the individual replication 1, an increment of "n" even by 20 blocks between 25 and 45 and by 25 blocks from 75 to 100 decreased the chi-square values from 7.28 to 4.04 and from 9.36 to 8.66, respectively. For the individual replication 2, the chi-square reduction pattern (from 11.56 to 4.86) was observed with every "n" increase by 25 blocks between 50 and 125. With the averaged, joint or pooled replications, the chi-square values generally increased with increased "n", however, with some fluctuation. Results from both Tables 2 and 3 showed that the Mack-Skillings test was relatively more sensitive to the differences (higher chi-square and lower corresponding *P* values) at all "n" sizes, compared to other methods of data analysis evaluated in this study.

Analyzing data from individual replications showed discordant null hypothesis test results at n = 25, 30, 40, 45, 100 and 125 all at $\alpha = 0.05$ (Table 3). As mentioned above, when increasing sample size for averaged, joint, and/or pooled replications, an immediate increase in a chi-square value was not always guaranteed. However, in all these replicated statistical alternatives, once a null hypothesis was rejected, a failure to reject it was not observed at a higher sample size, a characteristic not observed with the individual replications. This result (Table 3) showed that accounting for inter-panelist variation in duplicated ranking test can help improve not only discrimination capacity but also consistency in results of the null hypothesis testing, particularly when more panelists can be added to the analysis and the degree of differences in preference among samples is small (Table 3). Therefore, for samples that are similar (more variations in rank data from the two replications), a choice of data analysis methods is critical in order to derive valid conclusions.

Table 3.2 Comparisons of the chi-square values and *P* values across data analysis methods and sample sizes for the different-samples set.

	Averaged replication[a]		**Replication 1**		**Replication 2**		**Joint replication[b]**			**Pooled replication[c]**	
	Friedman's*		**Friedman's***		**Friedman's***		**Mack-Skillings****			**Friedman's***	
n	**X^2**	***P***	**X^2**	***P***	**X^2**	***P***	**X^2**	***P***	**n**	**X^2**	***P***
125	177.86	2.40E-39	177.74	2.50E-39	178.19	2.00E-39	406.53	5.30E-89	250	355.71	5.70E-78
100	142.82	9.70E-32	147.14	1.10E-32	138.66	7.80E-31	326.45	1.30E-71	200	285.64	9.40E-63
75	107.46	4.60E-24	109.95	1.30E-24	105.15	1.50E-23	245.62	4.60E-54	150	214.92	2.10E-47
50	71.04	3.70E-16	75.04	5.10E-17	67.36	2.40E-15	162.38	5.50E-36	100	142.08	1.40E-31
45	62.34	2.90E-14	65.38	6.40E-15	59.51	1.20E-13	142.5	1.10E-31	90	124.69	8.40E-28
40	56.71	4.80E-13	57.05	4.10E-13	56.45	5.50E-13	129.63	7.10E-29	80	113.43	2.30E-25
35	48.74	2.60E-11	50.8	9.30E-12	46.8	6.90E-11	111.41	6.40E-25	70	97.49	6.80E-22
30	45.87	1.10E-10	51.67	6.00E-12	40.47	1.60E-09	104.84	1.70E-23	60	91.73	1.20E-20
25	36.86	9.90E-09	42	7.60E-10	32.24	1.00E-07	84.25	5.10E-19	50	73.72	9.80E-17
10	13.95	0.0009	18.2	0.0001	10.4	0.0055	31.89	1.20E-07	20	27.9	8.70E-07

[a] Rank sums were obtained from the averaged rank data of each panelist from the two replications.

[b] Averaged rank sums were calculated as ($R_j^* = \sum_{i=1}^{n} r\left[\sum_{q=1}^{c} r_{ijq} / c\right]$), where c = 2. Such averaged rank scores are the sum of all scores from the *i*th block for the sample *j*, divided by c).

[c] Rank sums were obtained from the rank data of all panelists pooled from the two replications at certain "n" value to obtain 2*n = **n** blocks.

* Data were analyzed by the distribution-free Friedman test (1937).

** Data were analyzed by the method as described by Hollander and others (2013).

Table 3.3 Comparisons of the chi-square values and *P* values across data analysis methods and sample sizes for the similar-samples set

	Averaged replication[a] Friedman's*		**Replication 1 Friedman's***		**Replication 2 Friedman's***		**Joint replication[b] Mack-Skillings****			**Pooled replication[c] Friedman's***	
n	**X^2**	***P***	**X^2**	***P***	**X^2**	***P***	**X^2**	***P***	**n**	**X^2**	***P***
125	8.18	0.0168	12.35	0.0021	4.86	***0.0879***	18.69	8.70E-05	250	16.35	0.0003
100	6.95	0.031	8.66	0.0132	5.42	***0.0665***	15.87	0.0004	200	13.89	0.001
75	8.01	0.0183	9.36	0.0093	8.03	0.0181	18.3	0.0001	150	16.01	0.0003
50	8.32	0.0156	6.76	0.034	11.56	0.0031	19.02	0.0001	100	16.64	0.0002
45	5.03	***0.0807***	4.04	***0.1324***	7.64	0.0219	11.5	0.0032	90	10.07	0.0065
40	4.84	***0.089***	3.8	***0.1496***	6.65	0.036	11.06	0.004	80	9.68	0.0079
35	4.47	***0.1069***	4.51	***0.1046***	4.63	***0.0988***	10.22	0.006	70	8.94	0.0114
30	5.22	***0.0737***	6.47	0.0394	4.27	***0.1184***	11.92	0.0026	60	10.43	0.0054
25	5.18	***0.075***	7.28	0.0263	3.92	***0.1409***	11.84	0.0027	50	10.36	0.0056
20	2.93	***0.2317***	3.7	***0.1572***	3.1	***0.2122***	6.69	0.0353	40	5.85	***0.0537***
15	1.3	***0.522***	2.53	***0.2818***	0.93	***0.6271***	2.97	***0.2263***	30	2.6	***0.2725***
10	1.55	***0.4607***	2.6	***0.2725***	0.8	***0.6703***	3.54	***0.1701***	20	3.1	***0.2122***

[a] Rank sums were obtained from the averaged rank data of each panelist from the two replications.

[b] Averaged rank sums were calculated as ($R_j^* = \sum_{i=1}^{n} r\left[\sum_{q=1}^{c} r_{ijq}\,/c\right]$), where c = 2. Such averaged rank scores are the sum of all scores from the *i*th block for the sample *j*, divided by c).

[c] Rank sums were obtained from the rank data of all panelists pooled from the two replications at certain "n" value to obtain 2*n = **n** blocks.

* Data were analyzed by the distribution-free Friedman test (1937).

** Data were analyzed by the method as described by Hollander and others (2013).

Bold and italicized *P* values indicate acceptance of the null hypothesis (H_0: all samples are not different) at α= 0.05.

3.3.2 Method selection and sensitivity

When using individual replications for the null hypothesis testing, rejection of the null hypothesis was observed at every "n" size in both replications for the different-samples set with high degree of significance (Table 3.2). The highest observed *P* value was 0.0055 with only 10 panelists from the second replication (Table 3.2). In contrast, for the similar-samples set, definitive "n-based" cutoff of *P* values lower than 0.05 was not found in either replication. Analyzing data from individual replications showed discordant null hypothesis test results at various "n" sizes at $\alpha = 0.05$ (Table 3.3). This emphasized that analyzing data separately from individual replications the Friedman's test is not recommended.

Using an average of the rank sums from both replications in the Friedman's test accounted for inter-panelist variation; nonetheless, the *P* values obtained were higher than those in the Mack-Skillings test at every comparable "n" size (Table 3.3). Disregarding the between-blocks independence and pooling two replications into $\underline{\mathbf{n}} = nc = 2n$ blocks, naturally generated lower *P* values than averaging replications as the Friedman's test becomes less conservative when "n" increases relative to "k" (Boos and Stefansky 2013). However, the *P* values of pooling (converting) replications into blocks were not lower than the Mack-Skillings *P* values for all "n" sizes, implying that the Mack-Skillings test was relatively more sensitive to the difference.

At "n" $\geq$ 50, consistent conclusions (the null hypothesis was rejected) were observed among the three data analyses from averaged, joint, and/or pooled replications. However, the null hypothesis was rejected at a much lower "n" for joint and pooled replications (starting at 20-25), compared to that (n starting at 50) for the averaged replication. However, it was not the aim of this work to establish proper "n" size for duplicated multiple-samples ranking test, and more research is needed in this area.

3.3.3 Multiple comparison tests

For the different-samples set, all pairs (AB, AC and BC) of samples were found significantly different (data not shown). Table 3.4 shows the rank sum values obtained at various sample sizes between 10 and 125 panelists for the similar-samples set. It is important to remember that panelists were instructed to assign a score of "1" to the most preferred sample, "2" to the intermediate one, and "3" to the least preferred sample.

According to Table 3.4, the rank sum values are logical. Without exception across all "n" sizes, sample C (90% orange juice) had a higher rank sum score (i.e., tentatively less preferred) than samples A or B. The rank sums of sample A were mainly lower than those of sample B, with some exceptions. Following the global null hypothesis tests (Table 3.3), the post-hoc multiple comparison procedure was applied on the data arranged in the same structure as shown in Table 3.5.

When individual replications were analyzed, discordant conclusions not only for the global null hypothesis tests (Table 3.3) but also for the post-hoc multiple comparisons (Table 3.5) were observed. Specifically, at n = 50 and 75, there was an agreement in the global null hypothesis results in both replications (Table 3.3); however, a disagreement in the post-hoc multiple comparison results, i.e., the pairwise difference was observed for BC for replication 1 but for AC for replication 2 (Table 3.5). This re-emphasized that analyzing data separately from individual replications using the Friedman's test is not recommended. With the averaged replication, the pairwise differences (AC and/or BC) were observed only when "n" reached 50, a much higher "n" when compared with the joint and/or pooled replications.

Table 3.4 Rank sums* by sample for the similar-samples set

	Averaged replication[a]			Replication 1			Replication 2			Joint replication[b]				Pooled replication[c]		
n	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**n**	**A**	**B**	**C**
125	239	235	276	236	232	282	242	238	270	416	408	490	**250**	478	470	552
100	190	188.5	221.5	189	187	224	191	190	219	330	327	393	**200**	380	377	443
75	139.5	140.5	170	144	135	171	135	146	169	242	244	303	**150**	279	281	340
50	88	96	116	93	92	115	83	100	117	151	167	207	**100**	176	192	232
45	80.5	88	101.5	85	84	101	76	92	102	139	154	181	**90**	161	176	203
40	71	78.5	90.5	74	76	90	68	81	91	122	137	161	**80**	142	157	181
35	62	68.5	79.5	63	67	80	61	70	79	107	120	142	**70**	124	137	159
30	52	58.5	69.5	52	57	71	52	60	68	89	102	124	**60**	104	117	139
25	43.5	47.5	59	44	45	61	43	50	57	75	83	106	**50**	87	95	118
20	35.5	38.5	46	37	36	47	34	41	45	61	67	82	**40**	71	77	92
15	27.5	29	33.5	28	27	35	27	31	32	48	51	60	**30**	55	58	67
10	17.5	19.5	23	17	19	24	18	20	22	30	34	41	**20**	35	39	46

[a] Rank sums were obtained from the averaged rank data of each panelist from the two replications.

[b] Averaged rank sums were calculated as ($R_j^* = \sum_{i=1}^{n} r\left[\sum_{q=1}^{c} r_{ijq} / c\right]$), where c = 2. Such averaged rank scores are the sum of all scores from the *i*th block for the sample *j*, divided by c).

[c] Rank sums were obtained from the rank data of all panelists pooled from the two replications at certain "n" value to obtain 2*n = **n** blocks.

*A, B, and C are treatments and were ranked without ties (1 = most preferred and 3 = least preferred).

Table 3.5 Significantly different sample pairs based on the Tukey's HSD and/or Mack-Skillings tests for the similar-samples set[x]

		Averaged Replication by HSD*	Replication 1 by HSD*	Replication 2 by HSD*		Joint replication by Mack-Skillings			Pooled replication by HSD*	
n	**CV[†]**	**Pairs**	**Pairs**	**Pairs**	**n**	**CV**	**Pairs**	**n****	**CV**	**Pairs**
125	37.1	BC	AC,BC	--	125	49	AC,BC	250	52.4	AC,BC
100	33.2	BC	AC,BC	--	100	43.9	AC,BC	200	46.9	AC,BC
75	28.7	AC,BC	BC	AC	75	38	AC,BC	150	40.6	AC,BC
50	23.4	AC	BC	AC	50	31	AC,BC	100	33.2	AC,BC
45	22.2	--	--	AC	45	29.4	AC	90	31.4	AC
40	21	--	--	AC	40	27.7	AC	80	29.7	AC
35	19.6	--	--	--	35	25.9	AC	70	27.7	AC
30	18.2	--	AC	--	30	24	AC	60	25.7	AC
25	16.6	--	AC	--	25	21.9	AC,BC	50	23.4	AC
20	14.8	--	--	--	20	19.6	AC,BC	40	21	--
15	12.8	--	--	--	15	17	--	30	18.2	--
10	10.5	--	--	--	10	13.9	--	20	14.8	--

* HSD = Final rank sum pairs were analyzed with the distribution-free experiment wise multiple comparisons procedure.

** Rank sums were obtained from the rank data of all panelists pooled from the two replications at certain "n" value to obtain 2*n = **n** blocks.

[†] CV= Critical value for paired hypothesis rejection (df= k-1 = 2).

[x]A, B, and C are treatments and were ranked without ties (1 = most preferred and 3 = least preferred).

[y] -- Indicates no significant differences.

The post-hoc multiple comparison results for the joint replication analyzed by the Mack and Skillings test (1980) vs. the pooled replication analyzed by HSD showed a similar pattern (Table 3.5). Although the rank sums of sample A were mainly lower than those of sample B (with some exceptions, Table 3.4), significant differences between A and B were not observed at any "n" sizes. The pairwise differences (AC and/or BC) were observed at "n" between 25 and 125. Slight differences in results of both data analysis methods were observed at a lower "n" between 20 and 25, which may not be adequate to lead to a conclusion that the Mack-Skillings test are more sensitive to the differences. For a non-replicated ranking test, the more sensitive (to the difference) method would have lower critical values (CV) at a given k and "n". The Mack-Skillings multiple comparison tests utilize intermediate CV values between HSD and LSD-non-parametric paired-wise test; however, while allowing an experiment-wise multiple comparison test. For example, at "n" = 50 or **n** = 100, the CV values for pooled HSD, Mack-Skillings, and pooled LSD would be 33.2 (Lawless and Heyman 2010), 31, and 28 (Christensen and others 2006), respectively. In this study, when considering the CV values (Table 3.5), we can observe that analyzing the same N (= nkc) for the Mack-Skillings test or N (= **n**k) for a pooled HSD test, the former test required a lower CV value to analyze equal absolute rank sum differences, implying a more sensitive method.

3.3.4 Chi-square approximation and exact permutation distributions of the Mack-Skillings test

Up to this point, the Mack-Skillings chi-square approximation was used to calculate the *P* values of the test (Tables 3.2 and 3.3) to illustrate some advantages of using replications on multiple-samples ranking tests and some disadvantages of using the Friedman's test on replicated ranking, i.e., either losing power by only using the median rank scores of the replications or violating the assumption of independence between blocks. The Mack-Skillings *P* values can also be obtained from the exact permutation distribution of the test or a Monte Carlo simulation.

According to Bi (2009) both options are less conservative compared to the chi-square approximation of the Durbin statistic (a Friedman's test extension for incomplete block designs). Moreover, Hollander and others (2013) suggested using the exact test if the number of replications is $c < 4$, especially with a low significance level, e.g., $\alpha = 0.01$.

Table 3.6 Comparisons between the Mack-Skillings *P* values obtained by chi-square approximation and exact permutation distributions for the similar-samples set

n	Mack-Skillings (MS)statistic*	*P* value (MS $\geq \mathbf{X^2}$)	Exact *P* value	*P* value difference**
125	18.69	8.70E-05	0.0001	0.000013
100	15.87	0.0004	0.0002	-0.0002
75	18.3	0.0001	0.0002	0.0001
50	19.02	0.0001	1.00E-05	-0.00009
45	11.5	0.0032	0.0013	-0.0019
40	11.06	0.004	0.0022	-0.0018
35	10.22	0.006	0.0028	-0.0032
30	11.92	0.0026	0.0016	-0.001
25	11.84	0.0027	0.0014	-0.0013
20	6.69	0.0353	0.026	-0.0093
15	2.97	***0.2263***	**0.203**	-0.0233
10	3.54	***0.1701***	**0.1537**	-0.0164

* Both alternatives use the same computed Mack Skillings statistic.
** *P* value difference = Exact *P* - *P* (MS $\geq \mathbf{X^2}$); a negative number indicates a larger *P* value calculated from the chi-square approximation.
Bold and italicized *P* values indicate acceptance of the null hypothesis (H_0: all samples are not different) at $\alpha = 0.05$.

Comparisons between the Mack-Skillings *P* values obtained by chi-square approximation and exact permutation distributions for the similar-samples set is given in Table 3.6. The *P* values with the exact tests were generally lower (with a couple of exceptions) than those obtained using the chi-square approximation. Nevertheless, the largest difference in *P* values between the two methods was 0.023 at n = 15, and as n increased, the difference generally decreased. With the degree of differences between samples of the similar-samples set, using the chi-square

approximation did not affect the conclusions of the hypothesis testing in this current study. However, when possible, an exact test is advised since the degree of differences in preference among samples is generally unknown. The function "pMackSkil" of the R package "NSM3," yields *P* values based on the exact distribution of the Mack-Skillings statistic; although a Monte Carlo simulation can also be used if specified. An example of the R codes is shown in Figure 3.1.

#Title: Mack-Skillings option 1 (package: ASBIO)

```
# For each panelist, order data: (A1, A2, B1, B2, C1, C2)
panelist1<-c (3.5,1.5,5.5,5.5,1.5,3.5)
panelist2<-c (1.5,1.5,5.5,5.5,3.5,3.5)
panelist3<-c (1.5,1.5,5.5,5.5,3.5,3.5)
panelist4<-c (1.5,1.5,3.5,5.5,5.5,3.5)
panelist5<-c (5.5,1.5,3.5,3.5,1.5,5.5)
subject<-
cbind(panelist1,panelist2,panelist3,panelist4,panelist5)
sample<-c(rep(1,2),rep(2,2),rep(3,2))
library(asbio)
MS.test(subject, sample, reps=2)
```

#Title: Mack-Skillings option 2 (package: ASBIO)

```
#Create a named vector; for example: e1
# By panelist, order data: (A1, A2, B1, B2, C1, C2)
e1 <-c(3.5 ,1.5  ,5.5 ,5.5 ,1.5, 3.5 ,
            1.5 ,1.5, 5.5 ,5.5 ,3.5 ,3.5 ,
            1.5, 1.5 ,5.5 ,5.5, 3.5 ,3.5,
            1.5 ,1.5, 3.5 ,5.5, 5.5 ,3.5,
            5.5 ,1.5, 3.5 ,3.5 ,1.5, 5.5)
e1
#create a named matrix; for example: ex1
#specify  (c*k = 6) as the number of rows
ex1 = matrix(e1, nrow = 6)
ex1
library(asbio)
MS.test(ex1, c(1,1,2,2,3,3), reps=2)
```

Mack-Skillings option 3

```
#Use matrix ex1 from option 2
 MStest<-function(x,cc){
#x is a matrix with 2*k rows and n columns
#k is number of treatments; cc is replications
#n is number of panelists
n<-dim(x)[2]
k<-dim(x)[1]/cc
rs<-rep(0,k)
for(i in 1:k){rs[i]<-sum(x[(cc*i-cc+1):(cc*i),])/cc}
rr<-(sum(x[1:cc,])/cc)^2
for(i in 2:k){
rr<-rr+(sum(x[(cc*i-cc+1):(cc*i),])/cc)^2
}
s<-12*rr/(k*(k*n*cc+n))-3*(k*n*cc+n)
cr<-qchisq(0.95,k-1)
pv<-1-pchisq(s,k-1)
all<-round(c(s,cr,pv),4)
cat("M-S Statistic:", s,"\n")
cat("Critical value of alpha=0.05:",cr,"\n")
cat("p-value:",pv,"\n")
cat("Rank sum for treatments:",rs,"\n") all  }
 MStest(ex1,2)
```

Mack-Skillings on exact distribution (package NSM3)

```
#Use matrix ex1 from option 2
library(NSM3)
 x<-array(as.vector(ex1),
dim=c(2,3,5),dimnames=list(Rep=c(1,2),
Treat=c("A","B","C"),Panelist=c(1:5)))
 x
pMackSkil(x)
```

Critical values

```
#q(α/2, k) = q(0.05/2, 3)
cRangeNor(0.025,3)


#ms(α, k, n , c) = ms (0.05,3,5,2)
cMackSkil(0.05,3,5,2)
```

Figure 3.1 The R codes for Mack-Skillings global null hypothesis and multiple comparisons testing

3.4 Discussion

Collectively, results from Tables 3.2 and 3.3 suggested that, depending on the degree of differences among samples, a choice of data analysis methods may be very critical to derive valid and consistent conclusions at varying "n" sizes. For the similar-samples set, the most sensitive method was obtained when using the joint replication analyzed by the Mack-Skillings test, followed by the pooled replication analyzed by the Friedman's test (Table 3.3). The latter method took individual replications per panelists as independent blocks, a violation of the first assumption of the ranking test. The observed pattern that the Mack-Skillings test delivered lower *P* values than the Friedman's test that uses individual replications as independent blocks in this study is supported by Conover (1971) who described that the Friedman's test loses power when only k=3 samples are evaluated, while power is gained when "k" is increased. Analyzing data from individual replications showed discordant null hypothesis test results at various "n" sizes at $\alpha = 0.05$ (Table 3.3), hence should be avoided. Using an averaged rank sums improved agreement of conclusions; however, more panelists are required. As Hollander and others (2013) remarked, some information is lost when averaged or median (which equal averaged rank sums when $c = 2$) rank sums are used.

As previously discussed in the Materials and Methods section, instead of ranking c=2 replicates of "k" samples separately, one joint ranking of $ck = 2k$ presented samples (possibly called a double ranking test, or internally replicated ranking test) can be performed if the test does not carry sensory fatigue effect. For example, a visual preference ranking of $ck = 6$ or 8 total samples with different three-digit identification codes from three or four original treatments. A non-sensory research example given in Hollander and others (2013) using data from Campbell and Pelletier (1962) analyzed with the "R" software with the Mack-Skillings (1980) structured as: c =

3 replications, k = 4 treatments (laboratories), and n = 3 blocks (Niacin enrichment levels) can be found on the R package "Asbio" described by (Manly and others, 2015). Figure 3.1 shows two alternative R codes for the chi-square approximation of the test, one using the "Asbio" package and the other not requiring the package. Additionally, an exact test based on the distribution of the Mack-Skillings is shown in Figure 3.1, along with R codes for critical values for a multiple comparisons test. This test is less conservative than the chi-square approximation with less than four replications ($c < 4$), especially at a small α level.

Ongoing research is being performed to compare the effects of sample presentations (serving all 6 samples at once vs. serving 3 samples twice) and the complexity of the attributes evaluated (color vs. flavor) for both preference and intensity. Descriptions of the Mack-Skillings derivation, motivation and proportional frequencies theory are available (Mack and Skillings 1980; Hollander and Wolfe 2013).

In addition to the alternative non-parametric methods for handling replicated rank data as recommended by Hollander and others (2013) and demonstrated in this study, Boos and Stefanski (2013) advocated a weighed sum for the Wilcoxon rank sum or Kruskal Wallis statistics within blocks developed by Van Elteren (1959). The procedure is rather laborious and a multiple comparisons method following this approach was not provided. Boos and Stefanski (2013) also suggested that this method was better suited for a larger number of replications. Conover (1971) proposed a generalization to the Friedman's test for the case of $c > 1$ or in its nomenclature $m > 1$; however, multiple comparison procedures were not provided either. Replicated multiple-samples ranking tests were also reported in joint analyses with descriptive methods by ranked-scaling; alternatively, the replications were handled with ANOVA on the Friedman ranks (Pecore and others 2015). Panelists who performed both rating and ranking tests concluded that ranking could

be less boring and monotonous while they felt more confident in their responses for preference ranking, and ranking regarding willingness to buy food products (Hein and others, 2008).

3.5 Conclusion

The Friedman rank sum test is perhaps the most commonly used data analysis of non-replicated rank preference data. Replicated preference test may increasingly gain relevance since it increases the number of replications per sample and hence reduces cost of sensory testing. This study demonstrated analyses of duplicated rank preference data using the Friedman vs. the Mack-Skillings tests. In addition, the Mack-Skillings computation and hypothesis testing were illustrated using the R software for both chi-square approximation and exact distributions. When test samples are similar or confusable in their characteristics, hence more variations in rank data from the two replications, a choice of data analysis methods is critical in order to derive valid conclusions. Analyzing rank data separately by replication yielded inconsistent conclusions across sample sizes, and is not recommended. In this study, when the number of available panelists is reduced, replicated tests analyzed with the Mack-Skillings distribution-free method showed improved discrimination among samples relative to the Friedman test applied on data from averaged or pooled replications. This study demonstrated that the Mack-Skillings test, which takes into account the within-panelist variation, is more sensitive and appropriate for analyzing duplicated ranked data.

3.6 References

Anderson DA. 1988. Some models for overdispersed binomial data. Australian Journal of Statistics 30(2):125-48.

Basker D. 1988. Critical-vales of differences among rank sums for multiple comparisons. Food Technology 42(2):79-84.

Bi J. 2009. Computer-intensive methods for sensory data analysis, exemplified by rank test. Food Quality and Preference 20(3):195-202.

Bi J. 2006. Sensory Discrimination Tests and Measurements: Statistical Principles, Procedures and Tables. Blackwell Publishing, Ames, Iowa.

Boos DD, Stefanski LA. 2013. Permutation and Rank Tests (chapter 12). In Essential Statistical Inference (pp. 449-530). New York: Springer.

Bradley R, Kramer A. 1957. A quick, rank test for significance of differences in multiple comparisons. Food Technology 11(7):412.

Brockhoff PB. 2003. The statistical power of replications in difference tests. Food Quality and Preference 14(5-6):405-417.

Brockhoff PB, Schlich P. 1998. Handling replications in discrimination tests. Food Quality and Preference 9(5):303-12.

Campbell JA, Pelletier O. 1962. Determination of niacin (niacinamide) in cereal products. Journal of the Association of Agricultural Chemists 45(2):449-54.

Cochrane CYC, Dubnicka S, Loughin T. 2005. Comparison of methods for analyzing replicated preference tests. Journal of Sensory Studies 20(6):484-502.

Christensen ZT, Ogden LV, Dunn, ML, Eggett DL. 2006. Multiple Comparison Procedures for Analysis of Ranked Data. Journal of Food Science 71(2):S132-S143.

Conover WJ. 1971. Practical Nonparametric Statistics. John & Wiley and Sons, New York.

Edgington ES. 1980. Randomization tests. Dekker, New York.

Ennis DM, Bi J. 1998. The beta-binomial model: accounting for inter-trial variation in replicated difference and preference tests. Journal of Sensory Studies 13(4):389-412.

Friedman M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32(200):675-701.

Hein KA, Jaeger SR, Carr BT, Delahunty CM. 2008. Comparison of five common acceptance and preference methods. Food Quality and Preference 19(7):651-61.

Hollander M, Wolfe DA. 1973. Nonparametric Statistical Methods. John Wiley & Sons, New York.

Hollander M, Wolfe DA, Chicken E. 2013. Nonparametric Statistical Methods: New Jersey: John Wiley & Sons. 848 p.

Joanes D. 1985. On a rank sum test due to Kramer. Journal of Food Science 50(5):1442-4.

Kahan G, Cooper D, Papavasiliou A, Kramer A. 1973. Expanded tables for determining significance of differences for ranked data. Food Technology 27(5):61-9.

Kramer A. 1956. A quick, rank test for significance of differences in multiple comparisons. Food Technology 10(8):391-3.

Kramer A. 1960. A rapid method for determining significance of differences from rank sums. Food Technology 14(11):576-81.

Kramer A. 1963. Revised tables for determining significance of differences. Food Technology 17(2):1596.

Lawless HT, Heymann H. 2010. Sensory Evaluation of Food: Principles and Practices: Springer-Verlag, New York.

Mack GA, Skillings JH. 1980. A Friedman-type rank test for main effects in a two-factor ANOVA. Journal of the American Statistical Association 75(372):947-51.

Manly BF, McDonald LL, Thomas DL, McDonald TL, Erickson WP. 2015. 120 moose. Sel. Package 'asbio':120. Available from: http://up2date.hmdc.harvard.edu/yum-rep/CRAN/web/packages/asbio/asbio.pdf#page=120. Accessed 2016 January 14.

Meilgaard MC, Carr BT, Civille GV. 2006. Sensory Evaluation Techniques. CRC press, Florida.

Newell GJ, MacFarlane JD. 1987. Expanded tables for multiple comparison procedures in the analysis of ranked data. Journal of Food Science 52(6):1721-5.

Pecore S, Kamerud J, Holschuh N. 2015. Ranked-Scaling: A new descriptive panel approach for rating small differences when using anchored intensity scales. Food Quality and Preference 40:376-80.

Pitman EJG. 1936. Sufficient statistics and intrinsic accuracy. Mathematical Proceedings of the Cambridge Philosophical Society 32:567-79.

Rigdon EE. 1999. Using the Friedman method of ranks for model comparison in structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal 6(3):219-32.

Rinaman Jr WC. 1983. On distribution-free rank tests for two-way layouts. Journal of the American Statistical Association 78(383):655-9.

Schneider G, Chicken E, Becvarik R, Schneider MG. 2016. Package 'NSM3'.

Stone H, Sidel JL. 1993. Sensory Evaluation Practices. California: Academic Press, Inc. 338 p.

Van Elteren PH. 1959. On the combination of independent two sample tests of Wilcoxon. Bulletin of the International Statistical Institute 37:351-361.

CHAPTER 4. SERVING PROTOCOLS FOR DUPLICATED SENSORY RANKING TESTS: SINGLE SESSION VERSUS DOUBLE SERVING SESSIONS

4.1 Introduction

For sensory and consumer sciences, ranking tests can help determine differences in preference or intensity among multiple products (Lawless and Heyman, 2010). Preference ranking alone has shown more sensitivity to differences than hedonic liking in the elderly (Barylko-Pikielna and others, 2004), the general population (Villanueva and others, 2000), and children over four years old (Kimmel and others, 1994; Delarue and others, 2014).

In multiple-samples ranking tests, "n" panelists receive a set of "k" samples to rank according to an attribute intensity or personal preference. The data are ordinal, and non-parametric-tests and tables of critical values are widely chosen over traditional ANOVA. The tables of critical values are quick and easy to use for testing the null hypothesis; however, they fail to provide an estimation of the degree of differences (*P* value) between samples. Such tables have experienced constant evolution since the first set proposed by Kramer (1956). Later, Newell and MacFarlane (1987) and Basker (1988) used simulation of the maximum difference between all sets of paired comparisons to create critical values. Nonetheless, Christensen and others (2006) declared the method to be too conservative, and created new critical values for paired comparisons based on simulation of all possible paired differences in 10,000 simulated panels.

The Friedman (1937) rank-based nonparametric test, asymptotically follows a chi-squared distribution (Conover, 1999; Hollander and others., 2013), and the associated *P* value provides a measure of the degree of significance of the overall differences. Among non-parametric methods, the Friedman test (1937), detailed by Hollander and Wolfe (1973), is widely recommended for the randomized block design (RBD) without panelist*sample interaction used in sensory ranking (Joanes, 1985; Rayner and Best, 1990; Meilgaard and others, 2010, Lawless and Heyman, 2013).

Replicated sensory tests can potentially reduce the number of panelists, time and expenses, but it is critical to avoid compromising sensitivity. Special statistical tests for replicated studies not only account for inter-panelist variations but also adjust for the dependence of the responses within panelists. These adjustments limit the impact of disregarding the assumption or precondition of independence between blocks. Examples of statistics for replicated studies include an overdispersion-based model (Anderson, 1998), the beta-binomial test (Bi and Ennis 1988) and the corrected beta-binomial test (Brockhoff, 2003) for sensory discrimination (Bi, 2007).

For ranked data, the Mack and Skillings (M-S, 1980) test extended the test by Friedman (1937) to two or more replications within a block (without missing observations). This test avoids the misuse of replications from the same panelist as individual blocks, thus failing to achieve independence between blocks. Additionally, it yields the same results as Friedman (1937) for non-replicated data (Hollander and others, 2013). The procedure requires that all replications of the samples are ranked within a single block or panelist. For example, when duplicating a four-samples ranking, a panelist evaluates and ranks eight total physical samples if two replications are intended. Nonetheless, each ranking eight samples at once creates concerns of fatigue, adaptation, or memory interference.

The rank tests for the two-way layout avoid requiring normally distributed data; however, they are not free from assumptions. According to Conover (1971), “k” blocks should be independent, without the influence of a block over the scores of another. Most of the researches evaluating the impact of violation of the independence assumptions have focused on parametric ANOVA (Rigdon, 1999) and not methods such as the Friedman test. In Friedman-type tests, dependence can occur in two ways. If a panelist repeats tests and each replicate is considered a different block, then a violation of independence between blocks occurs. The other dependence,

between the scores provided by a single panelist (within block), does not represent a violation of the assumption because Friedman-type tests only require that all samples are ranked fully within a block. According to Hollander and others. (2013), the assumption is replaced by the requirement that under the null hypothesis, the results for the obtained rank sums come from equally likely individual scores. Such scores, in sensory evaluation, are obtained by fully ranking all samples from panelist 1 to k. According to Hollander and others (2013), the M-S test also replaces the assumption of independence between the samples within a panelist. The new assumption implies that all samples from every replication are ranked from 1 to c*k or vice versa within each block. With this originally ranked data or ranking transformation from continuous data, all scores configurations are equally likely under the null hypothesis.

In sensory ranking tests, especially in studies using the senses of taste and smell, ranking all c*k samples could implicate many samples, thus reducing discriminative efficiency due to fatigue, adaptation, memory interference, or memory decay. This problem can be accentuated when the number of samples or replications increases. As noted by Dacremont & Sauvageot (1997), the objective of replications in sensory testing is to make the maximum use of panelists until fatigue is detrimental to discrimination.

Increasing the number of samples in a taste ranking tests may reduce sensitivity due to saturation or memory problems (Valentin and others, 2012). According to O'Mahony (1986) and Meilgaard and others (2010) adaptation can negatively impact the sensitivity to all of the senses of panelists after exposure to repeated stimuli, affecting the efficiency of sensory tests. Besides adaptation, memory interference can reduce the sensitivity of tests with an increased number of samples (Lau and others, 2004). Furthermore, Meilgaard and others (2007) suggested that above three samples, taste ranking tests could lack the discriminative efficiency; however, other senses

such as vision have received less attention. The reduction in sensitivity in tests requiring a higher number of samples (3AFC vs. 2FC or triangle vs. the same-different test) has been documented for chemical senses due to fatigue or memory loss (Rousseau and O'Mahony, 1997; Rousseau and others, 1998; Dessirier and O'Mahony 1999; Rousseau and others., 1999). Although these studies emphasized the differences between two samples (discrimination tasting), the fundamentals could translate to three or more samples (ranking).

Color is an important attribute in influencing consumer perceptions of flavor (Spence, 2010), liking (Zellner and Durlach, 2003; Muggah and McSweeney, 2017), emotions (Gilbert and others, 2016), overall perception and purchase intent among other characteristics. Chambers and Wolfe (1996) suggested that ranking tests on visual evaluations might be less prone to fatigue than taste and aroma tests, possibly favoring a joint ranking session. Hence, the objective of this study was to examine the impact of applying the M-S test on duplicated visual ranked data served either with both replications in one serving session (1SS) or in separate replications (2SS) in experimental conditions. The ranking tests were performed on yellow color intensity.

4.2 Materials and methods

4.2.1 Study rationale

A previous study (Carabante and others, 2016) demonstrated that the M-S method can be used to analyze multiple-sample preference ranked data with two replications. It was also noted that accounting for intra-panelist information from two replicates with this method can potentially reduce the number of panelists required to detect differences in preference. The reduction in required panelists was especially important when samples were similar (confusable) rather than extremely different.

The aforementioned study required panelists to evaluate replications separately in two serving sessions (2SS) with a break period in between. Two separate replications were performed for one different and one similar sample set. The M-S analysis of separate replications required the 1, 2, and 3 scores from both replications within a set of k = 3 samples to be re-ranked into intermediate ranks for each panelist. Each of the six samples evaluated by a panelist obtained a score of either 1.5, 3.5, or 5.5, and each score appeared twice in every panelist. For example, a random panelist who scored 2, 1, and 3 for samples “A”, “B”, and “C” in the first ranking and 2, 3, and 1 in the second ranking of the same samples generates two groups of 1, 2, and 3 scores. The intermediate rank transformation (a new ranking of all six scores) yields scores of 3.5 and 3.5 for sample “A”, 1.5 and 5.5 for sample “B”, and 5.5 and 1.5 for sample “C”. Such intermediate ranks were then used in the M-S test.

On the other hand, instead of performing replications separately, panelists could receive both replications in a single multi-sample ranking session (1SS) and rank samples A1, A2, B1, B2, C1, and C2 in a counterbalanced design using six unique identification codes. Serving both replications at once to each panelist limits the dependency of scores within and between panelists, also avoiding observing ties that could reduce the power of the test. The separate replications alternative (2SS) eliminates dependency between blocks (panelists) but will always generate ties in the intermediate scores; nonetheless, with less influence than taking each replication as a separate block.

Hollander and others. (2013) stated that the M-S test can handle ties, but with a larger number of ties, power losses can occur. Obtaining data from a single joint ranking containing all replications evaluated once (1SS) will eliminate the ties. However, given the sizeable number of

samples to rank increments and a lack of rest period, fatigue, adaptation, and memory interference could play a more detrimental role in discrimination.

4.2.2 Data analysis of replicated ranked data with the M-S test

The M-S test (Mack and Skillings, 1980; Hollander and others, 2013) is the extension of the Friedman (Friedman 1937; Hollander and Wolfe, 1973) test for more than one replication per sample*panelist combination. As its non-replicated counterpart, it asymptotically follows a chi-squared distribution with degrees of freedom (df) = k-1. The computation of both tests requires the same parameters: n = a number of panelists and k = a number of samples; however, the M-S test includes the parameter "c" for a number of replications. The total number of rank scores or cells (in a matrix arrangement, where *i* denotes the *i*th panelist, *j* the j^{th} sample and *l* the *l*th replication) is now calculated by N = k*n*c. The test also uses rank sums; however, the new "weighted" rank sum are calculated as follows: ($R_j^* = \sum_{i=1}^{n} r\left[\sum_{l=1}^{c} r_{ijl}\,/c\right]$). This calculation requires that all of the sample*replication (k*c) combinations are ranked within a panelist. For example, panelist "i7", evaluating three samples in two replicates yields = k*c = 3*2 = 6 mutually dependent scores (from 1 to 6) from only three original samples. The calculation of the rank sums (R_j^*) adds all of the scores from a single (*j*th) sample regardless of the replication and then divides by the number of replications "c". Finally, The M-S computation is as follows:

$$\text{M-S} = \left(\frac{12}{k(N+n)}\right) \sum_{j=1}^{k} \left(R_j \frac{N+n}{2}\right)^2 = \left(\frac{12}{k(N+n)}\right) \left[\sum_{j=1}^{k} R_j^2\right] - 3(N+n)$$

The null hypothesis (Ho) stands: all k samples are not different (Ho: $R_1 = R_2 = \ldots = R_k$.)

To illustrate the computation of the test statistic (M-S), we analyzed the following example data set where three samples (k=3), replicated twice (c=2) by four panelists (n =4), produced the rank scores shown on the left half of Figure 4.1.

Obtained data (k=3, n=5, c=2)							Averaged rank data to accommodate ties						
n	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**	**n**	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**
1	3	2	2	3	1	1	1	5.5	3.5	3.5	5.5	1.5	1.5
2	2	2	3	3	1	1	2	3.5	3.5	5.5	5.5	1.5	1.5
3	2	2	3	3	1	1	3	3.5	3.5	5.5	5.5	1.5	1.5
4	1	3	2	2	3	1	4	1.5	5.5	3.5	3.5	5.5	1.5

A, B, and C are treatments. 1 and 2 indicates replication.

Figure 4.1 Example of averaged intermediate rankings from c = 2 replications, k = 3 samples, and n = 4 panelists for Mack-Skillings analysis

The right half shows the calculated joint rank scores from the two individual three-sample complete rankings from each panelist (intermediate ranks from two serving sessions). After obtaining the intermediate rank scores from the separate datasets, each weighted rank sum (R_j*) should be calculated for samples "A", "B", and "C":

R_A = (5.5 + 3.5 +3.5 +1.5 + 3.5 + 3.5 + 3.5 + 5.5) / 2 = 30/2 = 15

R_B = (3.5 + 5.5 + 5.5 + 3.5 + 5.5 + 5.5 + 5.5 + 3.5)/2 = 38/2 = 19

R_C = (1.5 + 1.5 + 1.5 + 5.5 + 1.5 + 1.5 + 1.5 + 1.5)/2 = 16/2 = 8

Note that all scores obtained from a sample are divided by the number of replications (c = 2).

Using the obtained weighted rank sums, we obtain the following M-S statistic:

$$\text{M-S}= \left(\frac{12}{3(24+4)}\right) [[15]^2 + [19]^2 + [8]^2] - 3(24 + 4) = 8.85$$

With degrees of freedom = k-1 = 2 and α = 0.05, the rejection critical value is 5.991; then, p (8.85 > 5.991) = 0.012. The null hypothesis (Ho: A=B=C) is rejected, showing that at least one paired comparison yielded significant differences.

The multiple comparisons procedure is also described by Hollander and Wolfe (2013), and its computation is as follows: $R_A - R_B \geq q_{\alpha,k} * \sqrt[2]{\frac{k\,(N+n)}{12}}$

Where $q_{\alpha,k}$ represents the α^{th} distribution percentile for all "k" sample independent and normal variables (Mack and Skillings, 1980). R_A and R_B represent the weighed rank sums of samples "A" and "B" from a sample set of "k" samples, evaluated by "n" panelists. This computation provides multiple comparisons based on experiment wise-error rates. Rinaman (1983) compared the asymptotic relative efficiency of the M-S test against several two-way layouts (including RBD designs) test alternatives, finding that it held the highest efficiency across several distributions. Therefore, he recommended the use of ranks even in scenarios in which the original datasets were not ranked data. Comparisons of the M-S test to alternatives exist for relatively large sample sizes originating in gene expression experiments with favorable results for many replications (Barrera and others, 2004). The M-S test also served as the platform for the rank test for multiple factors by Groggel and Skillings (1986).

4.2.3 Sensory study

A group of 75 panelists was recruited at the Louisiana State University Agricultural Center Campus in Baton Rouge, LA. To participate in the study, panelists should agree with and sign a consent form included in the research protocol approved (IRB # HE 15-9) by the Louisiana State University (LSU) Agricultural Center Institutional Review Board. Before their initial participation, panelists were screened according to the following criteria: availability for repeated visits, no allergies or adverse reactions to the ingredients in orange juice, and lack of known sensory deficits such as impaired vision or color blindness. Each panelist performed six yellow color intensity complete-multiple ranking tests (Table 4.1). The panelists evaluated two sample sets, including a similar sample set (100, 95 and 90% orange juice) and a different sample set (100, 70 and 40% orange juice). For each sample set, three ranking tests were performed, including two separate replications (2SS) and one ranking of six samples containing both replications (1SS). Panelists

were instructed to assign a score of "1" to the highest yellow color intensity and a 3 to the lowest when they ranked two replications separately in two serving sessions (2SS). When panelists ranked two joint replications in one serving session (1SS) a score of 6 was assigned to the lowest yellow color intensity. The model product used was 100% Minute Maid® orange juice (Minute Maid, Chicago, IL), without pulp. The panelists completed all tests within a period of three weeks and never performed more than two ranking tests per day, with at least 15 minutes of rest between the two tests.

Table 4.1 Multiple-sample ranking test sessions performed by each panelist

Degree of difference*	Ranking test**	k†	Percentage of orange juice per sample					
Similar Samples (Set 1)	Test 1	3	100			95		90
	Test 2	3	100			95		90
	Test 3	6	100	95	90	100	95	90
Different samples (Set 2)	Test 4	3	100			70		40
	Test 5	3	100			40		40
	Test 6	6	100	70	40	100	70	40

*Relative degree of yellow color divergence between samples of a single ranking test.
** Panelists completed the six tests in three weeks in a counter balanced arrangement.
Tests 1 and 2 are separate replicates of the similar sample set. Tests 4 and 5 are separate replicates for the different sample set. † Number of samples ranked per set.

In each visit, the panelists completed one serving session protocol of either from the different or similar sample set. The samples within a session, the serving protocols, and the sample sets were presented to the panelists in a counter-balanced system. Unique three-digit codes were assigned to each sample regardless of replication to avoid influence from previous tests performed by a panelist. The ranking sessions were performed in 15 partition booths equipped with cool natural white lights. The data were collected with the software (Compusense release 5.6, Compusense Inc., Guelph, Ontario, Canada).

4.2.4 Colorimetric analysis

A colorimetric analysis was performed to obtain a frame of reference about the magnitude of differences between samples and its relationship to the perceived differences in ranking alternatives. Color analysis was performed using a CIE-L*a*b* (McLaren, 1976) scaled Minolta colorimeter, model BC-10 (Minolta Co., Osaka, Japan). Eight individually prepared 25 mL aliquots of each sample served as experimental units (N=24). Each measurement was performed in 2-oz. soufflé cups in a sensory partition booth illuminated with the same white light that the panelists used. For each recording, the colorimeter lens (protected) was immersed approximately 3 mm in the orange juice to avoid a biasing headspace.

4.3 Results and discussion

4.3.1 Effect of serving protocol and method of analysis

For simplicity, the set of three samples composed of 100, 95 and 90% orange juice is denoted as the similar sample set. Likewise, the set of three samples composed of 100, 70 and 40% orange juice is donated as the different sample set. All analyses, follow an asymptotic chi-squared distribution with two degrees of freedom (df = k-1 = 3-1 = 2). With the same number of degrees of freedom, besides the comparisons of *P* values, chi-square statistics can be compared. All comparisons represent a significance level of 0.05 (α =0.05), but the trends apply to other significance levels (data not shown).

The obtained rank sums and the number of panelists for the similar and different sample sets are shown in Table 4.2. Visual appreciation revealed that the rank sums followed the expected pattern. In both sets, the rank sums were inversely proportional to the percentage of orange juice. For example, the samples with 100% orange juice showed the lowest rank sum (highest yellow color intensity), and the samples with 90% orange juice showed the highest rank sums (lowest

yellowness intensity). Also, the similar samples set showed relatively lower differences than the different sample set, as the range of the rank sums was narrower. The analysis of statistical methods revealed that both serving protocols using the M-S test had higher statistics than Friedman on median replications (Fr statistic, Table 4.3). Hollander and others (2013) suggested the last method as a conservative alternative for handling replications within non-parametric tests. This option accounts for between-panelists variation; however, it excludes the use of replication information (within-panelists). Higher test statistical values, either M-S or Friedman, reflect a greater degree of significance of differences between at least one pair of samples.

For the different sample sets (Table 4.3), all global null hypothesis tests (Ho: all orange juice samples are not different in yellow color intensity) yielded null hypothesis rejections with a high degree of significance. The lowest test statistic value was 18.20 for the Friedman test on the median of both replications at n = 10 panelists (p = 0.0001). However, at each given "n", both M-S variations showed much larger statistic values than the analysis using only their median. These differences, nevertheless, have relatively low importance compared to the similar sample set (Table 4.4). Thus, the median of the replications also showed high significant differences across "n" values in the different sample set. In the similar sample set (Table 4.4), from the two serving protocols of replicated rankings, based on the highest M-S statistical values, the ranking test using 2SS provided the highest yellow color discrimination. Except for n = 30; the M-S statistics were higher in the 2SS protocol than in the 1SS alternative (13 out of 14 total tests, with varying "n"). For example, at n = 30, the M-S statistics from 1SS was 42.47 ($p = 6x10^{-10}$), a slight but futile increase over that obtained from M-S on 2SS (41.27 and $p = 1.1x10^{-09}$). At all other "n" values, separating replications yielded higher discrimination between samples, also seen through the total differences between rank sums calculated at every n*method combination.

Table 4.2 Rank sums* by sample for the different and similar sample set

	Rank Sums Different Sample Set									***Rank Sums Similar Sample Set***								
		2SS[c]			**1SS**			**Median Replication[a]**			***2SS[c]***			***1SS[b]***			***Median Replication[a]***	
n	**1**	**0.7**	**0.4**	**1**	**0.7**	**0.4**	**1**	**07**	**0.4**	***1***	***0.95***	***0.9***	***1***	***95***	***0.9***	***1***	***0.95***	***0.9***
75	135.5	258.5	393.5	126	263	398.5	86.5	148	215.5	*194.5*	*278.5*	*314.5*	*205.5*	*284.5*	*297.5*	*116*	*158*	*176*
70	128	241	366	118.5	245.5	371	81.5	138	200.5	*177*	*261*	*297*	*191.5*	*265.5*	*278*	*106*	*148*	*166*
65	115.5	224.5	342.5	107	228	347.5	74	128.5	187.5	*164.5*	*240.5*	*277.5*	*173*	*248*	*261.5*	*98.5*	*136.5*	*155*
60	106	208	316	99.5	210.5	320	68	119	173	*152*	*221*	*257*	*158*	*229.5*	*242.5*	*91*	*125.5*	*143.5*
55	96.5	190.5	290.5	92	193	292.5	62	109	159	*139.5*	*202.5*	*235.5*	*145.5*	*211*	*221*	*83.5*	*115*	*131.5*
50	88	174	263	84.5	175.5	265	56.5	99.5	144	*123.5*	*183.5*	*207.5*	*130.5*	*194.5*	*200*	*75.5*	*105.5*	*119*
45	79.5	156.5	236.5	77	158	237.5	51	89.5	129.5	*113.5*	*166.5*	*192.5*	*112.5*	*176.5*	*183.5*	*68*	*94.5*	*107.5*
40	72	138	210	69.5	140.5	210	46	79	115	*98*	*151*	*171*	*101*	*157*	*162*	*59*	*85.5*	*95.5*
35	60.5	120.5	186.5	58	122	187.5	39	69	102	*79*	*133*	*145*	*82*	*140.5*	*145*	*49.5*	*76.5*	*84*
30	53	103	159	49.5	105.5	160	34	59	87	*68*	*116*	*131*	*66.5*	*122.5*	*126*	*41.5*	*65.5*	*73*
25	41.5	85.5	135.5	38	88	136.5	27	49	74	*53*	*93*	*106*	*57*	*103*	*102.5*	*34*	*54*	*62*
20	34	68	108	30.5	70.5	109	22	39	59	*44*	*78*	*88*	*46.5*	*83.5*	*80*	*27*	*44*	*49*
15	24.5	50.5	82.5	23	53	81.5	16	29	45	*34.5*	*55.5*	*67.5*	*38.5*	*58.5*	*60.5*	*21*	*31.5*	*37.5*
10	17	33	55	15.5	35.5	54	11	19	30	*22*	*38*	*45*	*22.5*	*40.5*	*42*	*13.5*	*21.5*	*25*

[a] Rank sums were obtained from the median rank data of each panelist from the two replications.

[b] For each panelist, one ranking session contained two replications (ranking 1 to k*c = 6). Rank sums were calculated as ($R_j^* = \sum_{i=1}^{n} r\left[\sum_{q=1}^{c} r_{ijq}\,/c\right]$), where c = 2.

[c] Each panelist completed both replications separately, and intermediate scores were calculated by re-ranking both replications within a panelist. Rank sums were calculated as ($R_j^* = \sum_{i=1}^{n} r\left[\sum_{q=1}^{c} r_{ijq}\,/c\right]$), where c = 2.

*1, 0.95, 0.9 are treatments indicating the proportion of orange used in the similar-samples set and were ranked without ties (1 = highest yellow color intensity 3 = least yellow color intensity).

Table 4.5 shows the rank sum differences and the total differences between all sample pairs by the protocols employed. Each "total" value represents the sum of the rank sum differences between the samples containing 100 vs. 95, 95 vs. 90, and 100 vs. 90% orange juice. Greater negative values (higher absolute values) represent greater separation between rank sums. With 2SS, the total differences were higher than in the 1SS protocol, supporting the conclusions based on the M-S statistics (except at n = 30). With 1SS, the total differences ranged from -39 at n = 10 to -184 at n = 75. Whereas, with 2SS, the range of the same "n" values was -46 to -240.

Table 4.3 Comparisons of the chi-square values (α =0.05) and *P* values across data analysis methods and sample sizes for the different samples set.

	2SS[b]			**1SS[b]**			**Median Replication[a]**	
	Mack-Skillings**			**Mack-Skillings****			**Friedman's***	
n	**MS Stat**	***P* > Chi2**	**Exact P**	**MS Stat**	***P* > Chi2**	**Exact P**	**Fr Stat**	***P* > Chi2**
75	253.76	7.88E-56	p<0.0001	282.88	3.74E-62	p<0.0001	111.02	7.80E-25
70	231.4	5.66E-51	p<0.0001	260.23	3.10E-57	p<0.0001	101.24	1.04E-22
65	226.62	6.17E-50	p<0.0001	254.25	6.18E-56	p<0.0001	99.15	2.96E-22
60	210.06	2.44E-46	p<0.0001	231.53	5.30E-51	p<0.0001	91.9	1.11E-20
55	195.57	3.40E-43	p<0.0001	208.84	4.48E-46	p<0.0001	85.56	2.63E-19
50	175.02	9.90E-39	p<0.0001	186.18	3.73E-41	p<0.0001	76.57	2.36E-17
45	156.52	1.03E-34	p<0.0001	163.56	3.04E-36	p<0.0001	68.48	1.35E-15
40	136.11	2.77E-30	p<0.0001	141.01	2.40E-31	p<0.0001	59.55	1.17E-13
35	129.7	6.86E-29	p<0.0001	136.91	1.87E-30	p<0.0001	56.74	4.77E-13
30	107.12	5.47E-24	p<0.0001	116.3	5.58E-26	p<0.0001	46.87	6.65E-11
25	101.12	1.10E-22	p<0.0001	110.89	8.32E-25	p<0.0001	44.24	2.47E-10
20	78.4	9.45E-18	p<0.0001	88.04	7.62E-20	p<0.0001	34.3	3.56E-08
15	64.3	1.09E-14	p<0.0001	65.2	6.95E-15	p<0.0001	28.13	7.78E-07
10	41.6	9.26E-10	p<0.0001	42.37	6.30E-10	p<0.0001	18.2	0.0001117

[a] Rank sums were obtained from the median rank data of each panelist from the two replications.
[b] Rank sums were calculated as ($R_j^* = \sum_{i=1}^{n} r\left[\sum_{q=1}^{c} r_{ijq} / c\right]$), where c = 2.
* Data were analyzed by the distribution-free Friedman test (1937).
** Source: Hollander and others (2013).

Table 4.4 Comparisons of the chi-square values (α =0.05) and *P* values across data analysis methods and sample sizes for the similar samples set.

	2SS[b]			1SS[b]			Median [a]	
	Mack-Skillings**			Mack-Skillings**			Friedman's*	
n	**MS Stat**	***P* > Chi2**	**Exact P**	**MS Stat**	***P* > Chi2**	**Exact P**	**Fr Stat**	***P* > Chi2**
75	57.78	2.84E-13	<0.0001	37.78	6.27E-09	<0.0001	25.28	3.24E-06
70	61.91	3.60E-14	<0.0001	35.69	1.78E-08	<0.0001	27.09	1.31E-06
65	58.36	2.13E-13	<0.0001	39.97	2.09E-09	<0.0001	25.53	2.86E-06
60	54.23	1.68E-12	<0.0001	39.43	2.74E-09	<0.0001	23.73	7.05E-06
55	49.43	1.84E-11	<0.0001	34.95	2.58E-08	<0.0001	21.63	2.01E-05
50	45.33	1.44E-10	<0.0001	34.12	3.90E-08	<0.0001	19.83	4.94E-05
45	41.17	1.15E-09	<0.0001	38.88	3.60E-09	<0.0001	18.01	0.0001227
40	40.67	1.48E-09	<0.0001	32.77	7.65E-08	<0.0001	17.79	0.0001372
35	43	4.59E-10	<0.0001	40.33	1.74E-09	<0.0001	18.81	8.21E-05
30	41.26	1.10E-09	<0.0001	42.47	6.00E-10	<0.0001	18.05	0.0001204
25	38.03	5.51E-09	<0.0001	31.9	1.18E-07	<0.0001	16.64	0.0002436
20	30.4	2.50E-07	<0.0001	23.84	6.65E-06	<0.0001	13.3	0.001294
15	21.26	2.42E-05	<0.0001	11.28	0.0035596	0.004	9.3	0.0095616
10	15.86	0.0003552	<0.0001	13.45	0.0011962	0.0012	6.95	0.0309618

[a] Rank sums were obtained from the median rank data of each panelist from the two replications.
[b] Rank sums were calculated as ($R_j^* = \sum_{i=1}^{n} r\left[\sum_{q=1}^{c} r_{ijq} / c\right]$), where c = 2.
* Data were analyzed by the distribution-free Friedman test (1937).
** Source: Hollander and others (2013).

Because the panelists were not instructed to evaluate a sample after another restricting a collective perspective of all samples in a ranking session (such as in taste), a panoramic view to rank samples provided an almost continuous reference for comparison between all samples. In this way, panelists avoided a complete interruption of each stimulus when comparing all samples. On this basis, memory interference or decay becomes a less relevant factor, contributing to the loss of sensitivity with a higher number of samples than in other proven attributes, e.g., taste (Rousseau and others, 2002; Lau and others, 2004).

Table 4.5 Rank sum differences and multiple paired comparison tests based on the Tukey's HSD and/or Mack-Skillings tests for the similar samples setx.

		2SS				1SS				Median Replication				
		by Mack-Skillings				by Mack-Skillings				by HSD*				
n	**CV****	**X1**	**X2**	**X3**	**Total**	**X1**	**X2**	**X3**	**Total**	**CV**	**X1**	**X2**	**X3**	**Total**
75	38	**-84**	-36	**-120**	-240	**-79**	-13	**-92**	-184	28.7	**-42**	-18	**-60**	-120
70	36.7	**-84**	-36	**-120**	-240	**-74**	-12.5	**-86.5**	-173	27.7	**-42**	-18	**-60**	-120
65	35.4	**-76**	**-37**	**-113**	-226	**-75**	-13.5	**-88.5**	-177	26.7	**-38**	-18.5	**-56.5**	-113
60	34	**-69**	**-36**	**-105**	-210	**-71.5**	-13	**-84.5**	-169	25.7	**-34.5**	-18	**-52.5**	-105
55	32.5	**-63**	**-33**	**-96**	-192	**-65.5**	-10	**-75.5**	-151	24.6	**-31.5**	-16.5	**-48**	-96
50	31	**-60**	-24	**-84**	-168	**-64**	-5.5	**-69.5**	-139	23.4	**-30**	-13.5	**-43.5**	-87
45	29.4	**-53**	-26	**-79**	-158	**-64**	-7	**-71**	-142	22.2	**-26.5**	-13	**-39.5**	-79
40	27.7	**-53**	-20	**-73**	-146	**-56**	-5	**-61**	-122	21	**-26.5**	-10	**-36.5**	-73
35	25.9	**-54**	-12	**-66**	-132	**-58.5**	-4.5	**-63**	-126	19.6	**-27**	-7.5	**-34.5**	-69
30	24	**-48**	-15	**-63**	-126	**-56**	-3.5	**-59.5**	-119	18.2	**-24**	-7.5	**-31.5**	-63
25	21.9	**-40**	-13	**-53**	-106	**-46**	0.5	**-45.5**	-91	16.6	**-20**	-8	**-28**	-56
20	19.6	**-34**	-10	**-44**	-88	**-37**	3.5	**-33.5**	-67	14.8	**-17**	-5	**-22**	-44
15	17	**-21**	-12	**-33**	-66	**-20**	-2	**-22**	-44	12.8	-10.5	-6	**-16.5**	-33
10	13.9	**-16**	-7	**-23**	-46	**-18**	-1.5	**-19.5**	-39	10.5	-8	-3.5	**-11.5**	-23

* HSD = Final rank sum pairs were analyzed with the distribution-free experiment-wise multiple comparisons procedure.
** CV= Critical value for paired hypothesis rejection (df= k-1 = 2).
X1 = R100- R95, X2 = R95-R90, X3= R100- 90, and were ranked without ties (1 = highest yellow color intensity and 3 = least yellow color intensity).
The bold values indicate pairwise significant at $\alpha = 0.05$
Total = X1 + X2 +X3 for each method.

Kinchla and Smyzer (1967) stated that the temporal continuity of visual stimuli reduces memory diffusion, aiding in discrimination. However, to the perceived wavelength reflected by the orange juice samples (yellow variations, among others), chromatic adaptation imposes a higher obstacle on sensitivity. Self-adaptation suggests that the perception of a stimulus is more difficult after the same stimulus was previously elicited (i.e., yellowness of the juice) than if the previously elicited stimulus was different (Rousseau and others, 1997; O'Mahony, 1986).

According to Fairchild (2013), repeated exposure of the retinal areas to energy reflecting a specific color reduces visual sensitivity. Moreover, evaluating six samples (1SS) takes longer than ranking three samples separately twice (2SS), especially with very similar samples, extending the exposure of the cones in the retina to the stimuli and increasing chromatic adaptation, which is a spatial- and time-dependent phenomena (Werner, 2014).

Ties from the intermediate rankings were not a relevant problem in reducing sensitivity, as the M-S on 2SS was less sensitive than the 1SS alternative only once in 14 tests (i.e., n = 30) for the similar-samples set (Table 4.4). The physiological sensitivity decrease from ranking duplicates in 1SS was greater than the impact of ties from intermediate rankings for an intensity test such as yellow color with highly similar samples. However, with extremely different samples, the 2SS were indeed less sensitive than 1SS.

There are no records of duplicated color rankings in the literature, but old records of color evaluations with ranking exist with panelists evaluating up to 10 samples at once. Nevertheless, the objective was measuring preference of green color intensity and not the intensities themselves (Buckle and Edwards, 1970). More recently, rankings have also been used to measure visual characteristics other than color, e.g., glossiness with six samples of coated Valencia oranges

(Contreras-Oliva, 2011). Also, overall appearance of raw beef steaks and fat appearance in raw beef steaks (Torrico, et al., 2014).

4.3.2 Effect of sample size on test statistics

When the number of panelists (n = blocks) is increased relative to the number of products or random variables, we expected a higher sensitivity to differences (Conover, 1990). To assess the influence of the number of panelists, we considered the change in test statistics (Mack-Skillings or Friedman) after adding five or 10 panelists. At each "n" value, all the results come from the exact same panelists. When differences exist, and are detected by the panelists, it is expected that adding more panelists will increase the significance of the differences. With a larger degree of differences between samples, reductions in test statistic values after adding panelist responses are also less likely given that less confusion yields lower variance.

With a different sample set (Table 4.3), each addition of only five panelists increased the significance of differences in every method. With similar samples (Table 4.4), a different behavior was observed. In both the M-S on 2SS and the median analyzed by Friedman, the only increases in panelists failing to produce a higher significance occurred from n = 35 to 40 and from 30 to 40, respectively. For example, in the test on 2SS at n = 35, the calculated M-S decreased from 43.0 to 40.67 after adding five panelists. Additionally, 15 more panelists were required to obtain a value higher than 43. Regardless, the M-S's statistic on data from the 1SS experienced several reductions after increases of five, 10, or more panelists. For example, the highest statistic appeared at n = 35 (MS = 42.47), and the highest number of panelists evaluated (75) produced a lower calculated statistic: 37.78. The highest number of reductions was in M-S's statistic on 1SS rather than 2SS, as evidenced by the increased difficulty of panelists to rank the six samples in the correct order

than in a three-sample ranking in which the variance was lower. Hence the 2SS protocol was more sensitive and consistent in hypothesis testing for the similar-samples set (Table 4.4)

4.3.3 P value estimates using the exact distributions of the M-S test

After obtaining the MS statistic, in addition to an asymptotic chi-square approximation, an exact *P* value can also be estimated based on the complete distribution of the M-S or a Monte-Carlo simulation using the package "NSM3" (Schneider and others., 2016); both computations can be obtained from the software R. The function pMackSkil yields an exact computation or a Monte Carlo simulation with more than 10,000 iterations if specified. According to Bi (2009) both approaches were less conservative than the chi-square asymptotic approximation for the Durbin–statistic, designed as an extension of the Friedman test for an incomplete block design. Hollander and others (2013) also recommend using an exact test with three or fewer replications per block and treatment combination.

Tables 4.3 and 4.4 show that the exact *P* values obtained from both the similar and the different sample sets often are slightly lower than those obtained by chi-square approximation (n =20-75), but *P* values below 0.0001 are not provided by the function. With the color intensities evaluated, the option for calculating the *P* value did not affect the null hypothesis conclusions; however, Hollander and others (2013) recommended using the exact P values if the number of replication is three or less. If possible, the exact approach should be used, given that, in most cases, the degree of product divergence is unknown, and with more confusable samples, the conclusions of the null hypothesis test could be affected. For R software commands, for replicated ranking scenarios, see Carabante and others (2016), where codes for global test statistics, multiple comparisons, and *P* value estimations on both approaches are available. Additionally, a description of other alternative analyses for replicated ranking is compared to the M-S test.

4.3.4 Multiple comparisons

All three possible paired comparisons (X1 = 100 vs. 95%, X2 = 95 vs. 90%, and X3 = 100 vs. 90% orange juice) were studied among all methods for the similar-samples set (Table 4.5). The Friedman test on the median of replications had unique critical values at each panel size obtained from the non-parametric HSD analog (experiment-wise multiple comparisons). The M-S test on 1SS and 2SS replications shared the same critical values obtained from the experiment-wise M-S multiple comparisons method. For each method, 42 possible paired differences were evaluated, given that each of the three paired comparisons was assessed at the 14 "n" possibilities (from 10 to 75 at every five-panelist increment).

With the different sample set (data not shown), all sample paired comparisons were significantly different, except for one. At n = 10, using the median of the separate replications, the samples with 100 and 70% orange juice showed a non-significant rank sums difference (diff = |-8|, < CV = |10.5|). The rest of the conclusions were unaffected by the protocol or the method selection, indicating less influence with a high degree of sample divergence.

The evaluation of significance ($\alpha = 0.05$) of the multiple comparisons method on similar samples is shown in Table 4.5. With similar samples, the protocol selection and the method showed higher influence in the number of significantly different pairs per n*method combination. This influence stems from the several contrasting conclusions obtained depending on the sample size. When comparing the two serving protocols analyzed with M-S and the multiple comparisons test, both tests yielded significant differences between 100 vs. 95% (X1) and 100 vs. 90% (X3), regardless of "n".

The Friedman test on the median of the replications failed to find a significant difference between 100 and 95% orange juice at n = 10 and 15, but with more panelists, it yielded the same

conclusions on X1 as the other two methods. The main contrast between 1SS and 2SS appeared in the comparison between 95 and 90% juice (X2). Neither the Friedman test on the median replications nor the M-S test on 1SS showed a single significant difference. Conversely, the M-S test on 2SS showed three significant differences, at n = 55, 60, and 65, whereas at n = 70 and 75, the differences were closer to statistical significance (Table 4.5).

Exploring the magnitude of the rank sum differences indicates that the test on 1SS showed lower rank sum differences for X3 and X2 than 2SS. With 2SS, X1 and X2, two sample pairs only differing in 5% orange juice, the rank sum differences achieved a higher balance than with joint rankings. The 1SS protocol tended to unbalance the differences towards X1 (100 vs. 95% orange juice), even if all the samples and methods were presented in a counter-balanced arrangement. Separating replications (2SS) also showed higher rank sum differences in X3, and the largest expected differences were with a 10% juice difference.

Although in both X1 and X2, the two samples only differed by 5% orange juice, a balanced linear difference may or may not necessarily represent the reality of the color difference perceived by consumers. Thus, the serving protocol more closely resembling the most accurate color difference perception of the panelist can be one showing balanced or unbalanced results between X1 and X2. This was considered not to punish the 1SS protocol for showing lower differences in X2 and allow the possibility that panelists found the difference harder to detect. Table 4.5 shows that joint rankings (ISS) produced less separation between 95 and 90% orange juice than serving replications separately with a break period (2SS), which showed fewer unbalanced rank sum differences for both pairs differing in 5% orange juice (X1, X2).

4.3.5 Instrumental colorimetric analysis

To investigate the differences between the samples and build a clear expectation of the magnitude of differences between samples (especially between X1 and X2), a colorimetric analysis conducted using a colorimeter based on a CIE-Lab scale is shown in Table 4.6. The Wilks' lambda test for differences between mean vectors showed significant differences between samples at the multivariate level ($P < 0.0001$). One-way ANOVA procedures showed significant differences for both the lightness (L*) and yellow/blue values (b*). In both parameters, all pairs of samples were significantly different based on a post hoc Tukey's test ($\alpha = 0.05$). Samples with less orange juice had lower lightness and less yellowness intensity. In yellowness intensity, the magnitude of the differences between 100-95 % and 95-90% orange juice showed a slightly higher difference for the first pair; however, the differences were relatively balanced (0.662 and 0.638, respectively). This balance indicated that if differences were found between 100 and 95% orange juice, findings showing differences between 95 and 90% orange juice was a plausible expectation. From the ranking data, the 1SS alternative could not reject the null hypothesis of no differences between 95 and 90% orange juice even with 75 panelists, while with the most balanced differences of the 2SS protocol; the difference between the pair in question were significant ($\alpha = 0.05$) despite requiring 55 panelists (Table 4.5). These results also showed higher efficiency for detecting expected significant differences in intensity rankings if replications were performed separately (2SS) with a break period (a break period of at least 15 min in this study).

Table 4.6 ANOVA and post hoc Tukey analysis of instrumental color data for the similar sample set.

% Orange juice	L*		a*		b*	
	F value	***P* >F**	**F value**	***P* >F**	**F value**	***P* >F**
	70.68	<.0001	1.41	0.267	62.89	<.0001
100	57.925 ± 0.104 ^{A}*		-1.2875 ± 0.099 ^{A}		13.725 ± 0.205 ^{A}	
95	57.5 ± 0.120 ^{B}		-1.225 ± 0.046 ^{A}		13.063 ± 0.130 ^{B}	
90	56.988 ± 0.223 ^{C}		-1.213 ± 0.125 ^{A}		12.425 ± 0.320 ^{C}	
Paired comparison	**L* Mean difference**		**a* Mean difference**		**b* Mean difference**	
100-95	0.425		-0.0625		0.662	
95-90	0.512		-0.012		0.638	
100-90	0.937		-0.0745		1.3	
Wilks' Lambda test for multivariate differences, F= 27.46					*P* > F	<0.0001

*Means with the same letter within a value (column) are not statistically different (*P* > 0.05)

4.4 Conclusions

This study showed that the M-S test was a suitable and efficient non-parametric analysis for replicated attribute intensity-ranked data. Regardless of the serving protocol of the replications, the M-S test showed higher discrimination than the median of individual replications analyzed with the Friedman test. The M-S test uses intra-block information to improve sensitivity to differences over averaging individual replications. A model study with two replications and three samples showed that when samples are relatively close in color intensity, separating the replications in complete individual ranking tests or serving sessions can help to prevent sensitivity loss due to fatigue or adaptation that is otherwise experienced when evaluating all replications together. These differences in discrimination were observed in both global null tests and multiple comparisons. When the samples of a set were extremely different, both serving protocols of replicated ranking performed with relatively similar discrimination efficiency.

4.5 References

Anderson DA. 1988. Some models for overdispersed binomial data. Australian Journal of Statistics 30(2):125-148.

Barrera L, Benner C, Tao Y-C, Winzeler E, Zhou Y. 2004. Leveraging two-way probe-level block design for identifying differential gene expression with high-density oligonucleotide arrays. BMC bioinformatics 5(1):42-55.

Barylko-Pikielna N, Matuszewska I, Jeruszka M, Kozlowska K, Brzozowska A, Roszkowski W. 2004. Discriminability and appropriateness of category scaling versus ranking methods to study sensory preferences in elderly. Food Quality and Preference 15(2):167-175.

Basker D. 1988. Critical-values of differences among rank sums for multiple comparisons. Food Technology 42(2):79-84.

Bi J. 2009. Computer-intensive methods for sensory data analysis, exemplified by Durbin's rank test. Food Quality and Preference 20(3):195-202.

Brockhoff PB. 2003. The statistical power of replications in difference tests. Food Quality and Preference 14(5):405-417.

Buckle K, Edwards R. 1970. Chlorophyll, colour and pH changes in HTST processed green pea puree. International Journal of Food Science & Technology 5(2): 173-186.

Chambers E IV, Wolf MB. 1996. Sensory Testing Methods, Pp. 46-50, 2nd ed. ASTM, West Conshohocken, PA: ASTM International.

Christensen ZT, Ogden LV, Dunn ML, Eggett DL. 2006. Multiple comparison procedures for analysis of ranked data. Journal of Food Science 71(2):32-43.

Conover W. 1971. Practical nonparametric statistics. John Wiley & Sons, Inc., New York.

Conover W. 1999. Practical nonparametric statistics. John Wiley & Sons, Inc., New York.

Contreras-Oliva A, Rojas-Argudo C, Pérez-Gago MB. 2011. Effect of solid content and composition of hydroxypropyl methylcellulose–lipid edible coatings on physicochemical, sensory and nutritional quality of 'Valencia' oranges. International Journal of Food Science & Technology 46(11):2437-2445.

Dacremont C, Sauvageot F. 1997. Are replicate evaluations of triangle tests during a session good practice? Food Quality and Preference 8(5–6):367-372.

Delarue J, Lawlor B, Rogeaux M. 2014. Rapid sensory profiling techniques: Applications in new product development and consumer research: Elsevier.

Dessirier JM, O'Mahony M. 1998. Comparison of d′ values for the 2-AFC (paired comparison) and 3-AFC discrimination methods: Thurstonian models, sequential sensitivity analysis and power. Food Quality and Preference 10(1):51-58.

Ennis DM, Bi J. 1998. The beta-binomial model: accounting for inter-trial variation in replicated difference and preference tests. Journal of Sensory Studies 13(4):389-412.

Fairchild MD. 2013. Color appearance models: John Wiley & Sons.

Friedman M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32(200):675-701.

Gilbert AN, Fridlund AJ, Lucchina LA. 2016. The color of emotion: A metric for implicit color associations. Food Quality and Preference 52:203-210.

Groggel DJ, Skillings JH. 1986. Distribution-free tests for main effects in multifactor designs. The American Statistician 40(2):99-102.

Hollander M, Wolfe DA. 1973. Nonparametric statistical methods. John Wiley & Sons

Hollander M, Wolfe DA, Chicken E. 2013. Nonparametric statistical methods: John Wiley & Sons.

Joanes D. 1985. On a rank sum test due to Kramer. Journal of Food Science 50(5):1442-1444.

Kimmel SA, Guinard J. 1994. Sensory testing with young children. Food Technology.

Kinchla R, Smyzer F. 1967. A diffusion model of perceptual memory. Perception & psychophysics 2(6):219-229.

Kramer A. 1956. A quick, rank test for significance of differences in multiple comparisons. Food Technology 10(8): 391-392.

Lau S, O'Mahony M, Rousseau B. 2004. Are three-sample tasks less sensitive than two-sample tasks? Memory effects in the testing of taste discrimination. Perception & psychophysics 66(3):464-474.

Lawless HT, Heymann H. 2010. Sensory evaluation of food: principles and practices: Springer Science & Business Media.

Mack GA, Skillings JH. 1980. A Friedman-type rank test for main effects in a two-factor ANOVA. Journal of the American Statistical Association 75(372):947-951.

McLaren K. 1976. XIII—The development of the CIE 1976 (L* a* b*) uniform colour space and colour-difference formula. Journal of the Society of Dyers and Colourists 92(9):338-341.

Meilgaard MC, Carr BT, Civille GV. 2006. Sensory evaluation techniques: CRC press.

Muggah EM, McSweeney MB. 2017. Females' attitude and preference for beer: a conjoint analysis study. International Journal of Food Science & Technology 52(3):808-816.

Newell G, MacFarlane J. 1987. Expanded tables for multiple comparison procedures in the analysis of ranked data. Journal of Food Science 52(6):1721-1725.

O'Mahony M. 1986. Sensory adaptation. Journal of Sensory Studies 1(3-4):237-258.

Rayner J, Best D. 1990. A comparison of some rank tests used in taste-testing. Journal of the Royal Society of New Zealand 20(3):269-272.

Rigdon EE. 1999. Using the Friedman method of ranks for model comparison in structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal 6(3):219-232.

Rinaman Jr WC. 1983. On distribution-free rank tests for two-way layouts. Journal of the American Statistical Association 78(383):655-659.

Rousseau B, Meyer A, O'Mahony M. 1998. Power and sensitivity of the same-different test: comparison with triangle and duo-trio methods. Journal of Sensory Studies 13(2):149-173.

Rousseau B, O'Mahony M. 1997. Sensory difference tests: Thurstonian and SSA predictions for vanilla flavored yogurts. Journal of Sensory Studies 12(2):127-146.

Rousseau B, Rogeaux M, O'Mahony M. 1999. Mustard discrimination by same–different and triangle tests: aspects of irritation, memory and τ criteria. Food Quality and Preference 10(3):173-184.

Rousseau B, Stroh S, O'Mahony M. 2002. Investigating more powerful discrimination tests with consumers: Effects of memory and response bias. Food Quality and Preference 13(1):39-45.

Schneider G, Chicken E, Becvarik R, Schneider MG. 2015. Package 'NSM3'.

Spence C, Levitan CA, Shankar MU, Zampini M. 2010. Does food color influence taste and flavor perception in humans? Chemosensory Perception 3(1):68-84.

Torrico DD, Jirangrat, W, Scaglia, G, Malekian F, Janes ME, McMillin KW, Prinyawiwatkul W. 2014. Proximate and fatty acid compositions and sensory acceptability of Hispanic consumers towards rib-eye steaks from forage-finished steers. International Journal of Food Science & Technology 49(8):1788-1798.

Valentin D, Chollet S, Lelievre M, Abdi H. 2012. Quick and dirty but still pretty good: A review of new descriptive methods in food science. International Journal of Food Science & Technology 47(8):1563-1578.

Villanueva ND, Petenate AJ, Da Silva MA. 2000. Performance of three affective methods and diagnosis of the ANOVA model. Food Quality and Preference 11(5):363-370.

Werner A. 2014. Spatial and temporal aspects of chromatic adaptation and their functional significance for colour constancy. Vision research 104:80-89.

Zellner DA, Durlach P. 2003. Effect of Color on Expected and Experienced Refreshment, Intensity, and Liking of Beverages. The American Journal of Psychology 116(4):633-647.

CHAPTER 5. SERVING DUPLICATES IN A SINGLE SESSION CAN SELECTIVELY IMPROVE THE SENSITIVITY OF DUPLICATED INTENSITY RANKING TESTS

5.1 Introduction

Because of their simplicity and sensitivity, sensory ranking tests with multiple samples are important preference and intensity difference tools (Meilgaard and others, 2006; Lawless and Heyman, 2010). With this method, "n" panelists have to rank a complete set of "k" samples according to the perceived intensity of a specific attribute or their overall preference. Each panelist generates an ordinal vector with dependent scores for each sample. Ranking tests are still widely used among the food industry with published applications for product screening (Bloom and Lee, 2016), preference (Mennella and others, 2017), and attribute difference (Urbanus and others, 2014); the later also involved the use of replications. The ordinal nature and dependency between the scores within a vector make the Friedman (1937) test a widely chosen statistical analysis for sensory ranked data (Joanes, 1985; Meilgaard and others, 2016; Lawless and Heyman, 2010). Other options for analysis include tables for critical values for rapid null hypothesis testing, either globally (Basker, 1988; Christensen and others, 2016) or for paired comparisons (Christensen and others 2016).

Successful sensory evaluation techniques require a high level of sensitivity to differences and efficiency with financial and human resources. Incorporating replications in the test is a viable option for optimizing resources when coupled with the appropriate statistical techniques that maximize effect information retrieval and restrict violations to independence. Stone and others (2012) recommended replications on sensory tests, making specific emphasis on duplications for increased power and control of within-panelist variations in discrimination testing rather than just increasing the number of panelists. Examples of replicated analysis, include an over dispersion

model (Anderson, 1998), the beta-binomial (Ennis and Bi, 1998) and corrected-beta binomial (Brockhoff, 2003) tests in discrimination testing. In ranking tests, the data from two blocks should be independent (Conover, 1971). The Mack-Skillings test (1980) extends the Friedman test to "c" > 1 replications, and controlling the dependency between the data from the same panelist (Hollander and others, 2013). The test requires that all samples are ranked within the same block, regardless of replication. Thus, a panelist ranking four samples in duplicates should generate a data vector with eight scores instead of two vectors with four scores.

Depending on the number of samples and replications intended, the physical serving of samples can be accomplished in a single session where a panelist ranks all samples at once; or in as many sessions as there are replications to reduce the possible adverse effects of a high number of samples. Several studies on the discrimination side of sensory testing suggested that a high number of samples is less desirable due to possible reduction in sensitivity to detect differences. The loss of sensitivity was mostly associated with adaptation, fatigue, and memory interference (O'Mahony, 1986; Rousseau and others, 2002; Lau and others 2004).

In duplicated ranking, separating the duplicates in two sessions appears as the first choice. Nevertheless, the Mack-Skillings test, requires that both data vectors are re-ranked into one, through intermediate rankings. With such re-ranking, ties between the data are unavoidable, possibly reducing the statistical power. Hollander and others (2013) stated that the M-S test can handle ties with relative efficiency; however, serving samples jointly not only eliminates ties, but also limits the dependency between the duplicates from a panelist. Increasing the number of samples in a test is not always undesirable, i.e., in the tetrad test when the noise imparted by an extra sample does not reduce statistical power (Ennis and Jesionka, 2011; Ennis, 2012).

With this study we tested two methods for sample serving in multiple samples ranking tests, and compared in two attributes dependent of different senses. Yellow color intensity and sweetness were evaluated in orange juice model sets varying in degree of difference between samples to examine if a joint ranking with duplicates served in the same session generates higher differences between samples or if it at least performs comparable to separating the replications into more than one session. Using orange juice models to evaluate color and sweetness instead of solutions adds dimensionality and a complexity level to the test (Bloom and Lee, 2016), and is better in resembling product testing. The results of this study are important for the development of new information in the use of duplicated ranking intensity tests.

5.2 Materials and methods

5.2.1 Panelists

A group of 75 panelists consisting of students, faculty and staff of the Louisiana State University were selected after successfully approving the following selection criteria: availability for repeated testing, lack of visual or taste impairment, e.g., color blindness or ageusia, overall health, sensory awareness and attitude. Before any participation, panelists had to agree with and sign a consent form as part of the research protocol approved by the Louisiana State University (LSU) Agricultural Center Institutional Review Board (IRB # HE 15-9).

5.2.2 Samples and sensory study

Two sample sets were designed with either similar or different samples. The similar sample set contained three samples with 100, 95 and 90% (w/w) orange juice. The different sample set contained three samples with 100, 70 and 40% (w/w) orange juice (Minute Maid®, Sugar Land, TX., U.S.A). Purified spring water was used to dilute the samples not containing 100% orange juice. All panelists separately evaluated both sample sets for both attributes (yellow color and

sweetness) using both protocols (1SS and 2SS) of duplicated ranking. One protocol required panelists to rank six samples jointly from 1 to 6 (1 = highest intensity) in one serving session (1SS), without knowing that there was another identical sample for each of the three concentrations. Thus, six three-digit random codes were used. With the two serving sessions protocol (2SS) panelists ranked the duplicates of a sample set separately (each one from 1 to 3, where 1 is the highest intensity), with a 10-minute break period. Six different blinding codes were used to discourage the idea of duplicates among panelists. To complete both protocols, a panelist had to perform three ranking tests or sessions for each sample set (similar and different sample sets). Both yellow color intensity and sweetness were evaluated separately, totaling 12 ranking sessions per panelists. The samples within a session, the sessions within a set, and the sample sets evaluated were presented in a counter balanced system, to reduce the influence of physiological and psychological effects produced by the presentation order. Because retasting or repeated color evaluation was allowed for confirmatory information, the counterbalanced system only applied to the first complete evaluation when it pertained to the samples in a set (Xia and others, 2016). The data were collected over a period of six weeks to fit the schedules of the participants; who never performed more than three sessions per day. The study was performed at the Sensory Services laboratory of the Louisiana State University Agricultural Center. The tests were performed using 15 partition booths, equipped with the software Compusense 5, release 5.6. (Compusense Inc., Guelph, ON, Canada). The booths were illuminated with clear natural lights for color analysis and red lights for sweetness.

5.2.3 Data analysis and the Mack-Skillings test.

For each sample set and attribute two types of data were collected. For the 1SS protocol, each of the 75 panelists generated a data vector of six mutually dependent scores (from 1 to 6).

With the 2SS protocol, the data came from two vectors of length = 3, containing 1's, 2's, 3's scores. For the 2SS data type, the Mack-Skillings (M-S) test, requires all data from a block (panelist) to be ranked jointly in one single vector for that block. With the 1SS protocol, the data fitted that requirement since collection, but with the 2SS protocol, intermediate ranking scores from re-ranking the scores from both replications were calculated. Re-ranking the two score vectors from the separate duplicates of a set evaluated by one panelist gives a score of 1.5 to each of the two "1" scores, a score of 3.5 to each of the two "2" scores and finally assigns "5.5" to the two "3" scores. The 75 panelists were randomized to obtain a new order from 1 to 75. After confirming every vector had six scores from the same panelists, the M-S test was applied to test the null hypothesis (Ho: There are no differences among samples) at every five-panelist increment from 10 to 75 panelists. At each increase in "n", the same panelists from the previous test were kept and only five new blocks were added. At a specific "n" value, the data for every attribute, set, protocol or session came from the same panelists. Additionally, multiple paired comparison tests were performed with the M-S multiple comparisons procedure, at all "n" values.

The M-S test is an extension of the Friedman Test for a randomized block design without treatment*block interaction for c > 1 replications. The *P* values based on the M-S statistic can be estimated from either a chi-squared approximation with degrees of freedom = k-1, where k is the number of treatments or samples. Also from an exact test or a Monte Carlo simulation where N panels of a size (n, k, c) are simulated; then the probability likelihood of such M-S statistic value is assessed based on its magnitude compared to the distribution of the simulated data. The computation of the M-S statistic follows:

$$\text{M-S} = \left(\frac{12}{k(N+n)}\right) \sum_{j=1}^{k} \left(R_j^* \frac{N+n}{2}\right)^2 = \left(\frac{12}{k(N+n)}\right) \left[\sum_{j=1}^{k} R_j^2\right] - 3(N+n)$$

The null hypothesis (Ho) stands: all k samples are not different (Ho: $R_1 = R_2 = ... = R_k$.); "k" represents the number of samples, "n" is the number of panelists, "c" are the number of complete replications. The total number of rank scores is N= k*n*c. R*j represents the weighted rank sum from the jth sample; calculated by adding all the scores of a sample from all replications, then dividing it by the number of replications ($R_j^* = \sum_{i=1}^{n} r\left[\sum_{l=1}^{c} r_{ijl}\,/c\right]$). Hollander and others (2013) and Mack and Skillings (1980) also provide an experiment-wise multiple comparisons non parametric procedure described by:

$$R_A - R_B \geq q_{\alpha,k} * \sqrt[2]{\frac{k\,(N+n)}{12}}$$

Where $q_{\alpha,k}$ is the α^{th} distribution percentile for all "k" sample independent and normal random variables (Mack and Skillings, 1980). R_A and R_B represent the weighed rank sums of samples "A" and "B" from a set of "k" samples. For application examples, refer to Hollander and others (2013) and Carabante and others (2016). The M-S statistics and *P* values were estimated using a Monte Carlo simulation with 10,000 iterations using the R software. Code alternatives can be found in a previous duplicated ranking introductory article (Carabante and others, 2016). These codes are similar in nature to those by Bi (2009) for the Durbin's tests for incomplete block designs.

5.3 Results

5.3.1 Measured rank sums

Table 5.1 shows the weighted rank sums for both attributes and both serving protocols. For the 2SS protocols, the two data vectors from each panelist were re-ranked into one block through intermediate rankings. With the 1SS protocol, the scores in the rank sums were the original scores of the data vectors provided by each panelist. .

Table 5.1 Rank sums by sample for the similar and different sample sets.

Set	n	Yellow Color 2SS			Sweetness 2SS			Yellow Color 1SS			Sweetness 1SS		
		100	**95**	**90**	**100**	**95**	**90**	**100**	**95**	**90**	**100**	**95**	**90**
Similar Sample Set	75	194.5	278.5	314.5	228.5	251.5	307.5	205.5	284.5	297.5	213	264.5	310
	70	177	261	297	215	237	283	191.5	265.5	278	201	245.5	288.5
	65	164.5	240.5	277.5	200.5	221.5	260.5	173	248	261.5	188.5	228.5	265.5
	60	152	221	257	182	204	244	158	229.5	242.5	172.5	210.5	247
	55	139.5	202.5	235.5	166.5	186.5	224.5	145.5	211	221	154	194.5	229
	50	123.5	183.5	207.5	147	174	204	130.5	194.5	200	138	179	208
	45	113.5	166.5	192.5	132.5	159.5	180.5	112.5	176.5	183.5	124	162.5	186
	40	98	151	171	117	139	164	101	157	162	111.5	142.5	166
	35	79	133	145	100.5	120.5	146.5	82	140.5	145	95.5	125	147
	30	68	116	131	83	102	130	66.5	122.5	126	79.5	111	124.5
	25	53	93	106	74.5	83.5	104.5	57	103	102.5	65	91	106.5
	20	44	78	88	59	70	81	46.5	83.5	80	50	71	89
	15	34.5	55.5	67.5	41.5	56.5	59.5	38.5	58.5	60.5	40	52.5	65
	10	22	38	45	24	37	44	22.5	40.5	42	26.5	36	42.5
Set	**n**	**100**	**70**	**40**	**100**	**70**	**40**	**100**	**70**	**40**	**100**	**70**	**40**
Different Sample Set	75	135.5	258.5	393.5	127.5	256.5	403.5	126	263	398.5	123.5	260	404
	70	128	241	366	118	241	376	118.5	245.5	371	115.5	243	376.5
	65	115.5	224.5	342.5	110.5	223.5	348.5	107	228	347.5	107	225.5	350
	60	106	208	316	103	206	321	99.5	210.5	320	99.5	208	322.5
	55	96.5	190.5	290.5	95.5	187.5	294.5	92	193	292.5	92	190.5	295
	50	88	174	263	88	170	267	84.5	175.5	265	84.5	173	267.5
	45	79.5	156.5	236.5	79.5	152.5	240.5	77	158	237.5	76	156.5	240
	40	72	138	210	70	137	213	69.5	140.5	210	68.5	139	212.5
	35	60.5	120.5	186.5	61.5	120.5	185.5	58	122	187.5	57	124.5	186
	30	53	103	159	53	104	158	49.5	105.5	160	49.5	107	158.5
	25	41.5	85.5	135.5	45.5	86.5	130.5	38	88	136.5	42	89.5	131
	20	34	68	108	38	69	103	30.5	70.5	109	34.5	72	103.5
	15	24.5	50.5	82.5	25.5	52.5	79.5	23	53	81.5	22.5	55	80
	10	17	33	55	18	35	52	15.5	35.5	54	15	37.5	52.5

*Indicates orange juice % in the samples. ** Rank values: 1 highest intensity, 3 = lowest intensit

Visual observation of the rank sums indicated that the data followed the specific logical expectations required to continue the study. When the degree of difference between samples was larger, the samples with 100 % orange juice obtained lower rank sums than in the similar sample set at the exact same "n", attribute and protocol (Table 5.1). For example, at n= 30, for sweetness ranking using the 2SS protocol, the rank sums for 100% orange juice were 53 for the different sample set and 83 for the similar sample set, given the wider spread of scores in the different sample set. This affirmed that with the similar sample set, the stimuli were more confusable. In every protocol, number of panelists, attribute and degree of difference, the sample with 100% orange juice obtained the lowest rank sum values, while the sample with less orange juice (90 or 40%) had the highest values. For example, at n = 40 for yellow color intensity with the 1SS protocol in the similar sample set, the rank sum for 100% orange juice was 101, a lower value than 157 (95%) and 162 (90%). This indicated that although the samples of the similar set were more confusable, in general differences could still be perceived. The following two subsections detail the measured degree of difficulty between attributes to assess which one was more difficult to rank and the effect of the serving protocols in relationship with the task difficulty

5.3.2 Evaluation of stimulus difficulty

A measure of the difficulty of correctly ranking the three samples of orange juice for either color or sweetness can be achieved comparing the M-S statistics (Table 5.2 for similar and Table 5.4 for different samples) or the total sum of paired rank sum differences (Table 5.3 for similar and Table 5.5 for different samples). Higher M-S values associate with larger overall differences between the samples in all ranking protocols. Tables 5.3 and 5.5 show all the M-S rank sum differences and their sum; where, X1 represents 100 – 95 % orange juice; X2 is 95 – 90 % and X3 is 100 - 90% orange juice. The *P* values associated with the Mack-Skillings statistic were obtained

from a Monte Carlo simulation with 10,000 iterations. This method or an exact test were recommended over a Chi Squared approximation for less than four replications (Hollander and others, 2013).

When the degree of difference between samples was lower (similar samples, Table 5.2), the yellow color ranking had higher M-S statistic values than sweetness, in 26 out of 28 comparisons (14 comparisons per protocol). Thus, panelists were more efficient in correctly ranking samples for yellow color than for sweetness. A similar conclusion about the higher complexity and degree of difficulty of sweetness can be obtained from the tables of rank sum differences (Table 5.3). Except for a few cases, the sum of the three rank sum differences at each "n" was higher in color than in sweetness (Table 5.3). The few exceptions where sweetness showed higher sum of differences than color occurred in the 1SS protocol (n = 75, 70, 50, 20, 15). Only one of these higher sum of differences made the M-S statistic higher for sweetness than color (n = 15). This could be explained by the more homogenous size of individual paired differences in sweetness with 1SS; especially with higher number of panelists. For example, at n = 70, with the 1SS protocol, the total difference in sweetness was: -175 compared to -173 in color, but in color, the difference between 100 and 95% orange juice was much higher than in sweetness (X1 color = -74 vs. X1 sweetness = -44.5). Adding this high difference to the large difference found in X3 (100 – 90% orange juice) increases the M-S stat for color, compared to the more homogenous differences in sweetness. With the 2SS protocol, the total differences were always higher in color than in sweetness; nevertheless, the comparison between protocols is further discussed in the next section. With both comparisons (M-S statistic or the sum of total differences) ranking of yellow color intensity showed less complexity (i.e., more sensitivity) than ranking of sweetness regardless of protocol; although, the global null hypothesis was rejected at every "n" value for both attributes.

Table 5.2 Comparison of Mack-Skillings statistics across serving protocols and attributes for the similar-sample set

n	Two Serving Sessions (2SS)				One Serving Session (1SS)			
	Color		Sweetness		Color		Sweetness	
	*M-S Sta**t*	*Exact P**	*M-S Stat*	*Exact P*	*M-S Stat*	*Exact P*	*M-S Stat*	*Exact P*
75	57.8	<0.0001	25.2	<0.0001	37.8	<0.0001	35.9	<0.0001
70	61.9	<0.0001	19.7	<0.0001	35.7	<0.0001	31.3	<0.0001
65	58.4	<0.0001	16.3	1.00E-04	40	<0.0001	26.1	<0.0001
60	54.2	<0.0001	18.8	<0.0001	39.4	<0.0001	26.4	<0.0001
55	49.4	<0.0001	18	1.00E-04	34.9	<0.0001	29.3	<0.0001
50	45.3	<0.0001	18.6	<0.0001	34.1	<0.0001	28.3	<0.0001
45	41.2	<0.0001	14.7	4.00E-04	38.9	<0.0001	24.9	<0.0001
40	40.7	<0.0001	15.8	2.00E-04	32.8	<0.0001	21.4	<0.0001
35	43	<0.0001	17.4	1.00E-04	40.3	<0.0001	21.8	<0.0001
30	41.3	<0.0001	21.3	<0.0001	42.5	<0.0001	20.3	<0.0001
25	38	<0.0001	10.8	0.0031	31.9	<0.0001	20.1	<0.0001
20	30.4	<0.0001	6.9	0.0259	23.8	<0.0001	21.8	<0.0001
15	21.3	<0.0001	7.1	0.0191	11.3	0.004	11.9	0.0018
10	15.9	<0.0001	11.8	0.0013	13.5	0.0012	7.4	0.0245

*Exact *P* values were calculated using a Monte Carlo procedure with 10000 iterations. At each "n" value, the data for each protocol and attribute came from the exact same panelists.

** The weighted rank sums used in the M-S statistic were calculated as: $R_j^* = \sum_{l=1}^{n} r\left[\sum_{l=1}^{c} r_{ijl} / c\right]$. The computation involves the sum of all the scores for the *j*th sample, then divided by "c". With duplicates, c= 2.

Table 5.3 Multiple comparisons test including weighted rank sum differences and total differences across serving protocols and attributes for the similar-sample set

		Two Serving Sessions 2SS				**One Serving Session 1SS**			
				Color					
n	*CV*	*x1*	*x2*	*x3*	*Sum*	*x1*	*x2*	*x3*	*Sum*
75	38.0	**-84**	-36	**-120**	-240	**-79**	-13	**-92**	-184
70	36.7	**-84**	-36	**-120**	-240	**-74**	-12.5	**-86.5**	-173
65	35.4	**-76**	**-37**	**-113**	-226	**-75**	-13.5	**-88.5**	-177
60	34.0	**-69**	**-36**	**-105**	-210	**-71.5**	-13	**-84.5**	-169
55	32.5	**-63**	**-33**	**-96**	-192	**-65.5**	-10	**-75.5**	-151
50	31.0	**-60**	-24	**-84**	-168	**-64**	-5.5	**-69.5**	-139
45	29.4	**-53**	-26	**-79**	-158	**-64**	-7	**-71**	-142
40	27.7	**-53**	-20	**-73**	-146	**-56**	-5	**-61**	-122
35	25.9	**-54**	-12	**-66**	-132	**-58.5**	-4.5	**-63**	-126
30	24.0	**-48**	-15	**-63**	-126	**-56**	-3.5	**-59.5**	-119
25	21.9	**-40**	-13	**-53**	-106	**-46**	0.5	**-45.5**	-91
20	19.6	**-34**	-10	**-44**	-88	**-37**	3.5	**-33.5**	-67
15	17.0	**-21**	-12	**-33**	-66	**-20**	-2	**-22**	-44
10	13.9	**-16**	-7	**-23**	-46	**-18**	-1.5	**-19.5**	-39
				Sweetness					
n	*CV*	*x1*	*x2*	*x3*	*Sum*	*x1*	*x2*	*x3*	*Sum*
75	38.0	-23	**-56**	**-79**	-158	**-51.5**	**-45.5**	**-97**	-194
70	36.7	-22	**-46**	**-68**	-136	**-44.5**	**-43**	**-87.5**	-175
65	35.4	-21	**-39**	**-60**	-120	**-40**	**-37**	**-77**	-154
60	34.0	-22	**-40**	**-62**	-124	**-38**	**-36.5**	**-74.5**	-149
55	32.5	-20	**-38**	**-58**	-116	**-40.5**	**-34.5**	**-75**	-150
50	31.0	-27	-30	**-57**	-114	**-41**	-29	**-70**	-140
45	29.4	-27	-21	**-48**	-96	**-38.5**	-23.5	**-62**	-124
40	27.7	-22	-25	**-47**	-94	**-31**	-23.5	**-54.5**	-109
35	25.9	-20	**-26**	**-46**	-92	**-29.5**	-22	**-51.5**	-103
30	24.0	-19	**-28**	**-47**	-94	**-31.5**	-13.5	**-45**	-90
25	21.9	-9	-21	**-30**	-60	**-26**	-15.5	**-41.5**	-83
20	19.6	-11	-11	**-22**	-44	**-21**	-18	**-39**	-78
15	17.0	-15	-3	**-18**	-36	-12.5	-12.5	**-25**	-50
10	13.9	-13	-7	**-20**	-40	-9.5	-6.5	**-16**	-32

**Bolded fonts represent a significant paired difference at α = 0.05. X1 = R (100%)-R (95%), X2 = R (95%)-R(90%), X3 = R(100%)-R(90%).

*The weighted rank sums used in the M-S statistic were calculated as: $R_j^* = \sum_{i=1}^{n} r\left[\sum_{l=1}^{c} r_{ijl}\,/c\right]$. The computation involves the sum of all the scores for the *j*th sample, then divided by "c". With duplicates, c= 2

With the different samples set, the panelists did not experience problems ranking the intensities in the correct order for color and sweetness. With such degree of difference, the complexity of the attributes was barely different numerically and nonexistent for practical terms. In general, sweetness was a harder attribute to correctly rank than yellow color intensity, due to higher complexity, especially, and more importantly with similar samples. The higher difficulty for sweetness with similar samples was reduced with different sample sets, where both attributes had very high and relatively similar M-S values (Table 5.4) and total differences (Table 5.5) due to less variation between rankings.

Table 5.4 Comparison of Mack-Skillings statistics across serving protocols and attributes for the different-sample set*

n	**Two Serving Sessions 2SS****		**One Serving Session 1SS**	
	Color	Sweetness	Color	Sweetness
75	253.8	290.6	282.9	299.8
70	231.4	271.9	260.2	278.1
65	226.6	249.2	254.2	259.6
60	210.1	226.5	231.5	236.9
55	195.6	206.1	208.8	214.1
50	175.0	183.5	186.2	191.4
45	156.5	165.1	163.6	170.8
40	136.1	146.3	141.0	148.1
35	129.7	125.6	136.9	135.9
30	107.1	105.0	116.3	113.3
25	101.1	82.6	110.9	90.7
20	78.4	60.4	88.0	68.2
15	64.3	55.5	65.2	63.3
10	41.6	33.0	42.4	40.7

* The weighted rank sums used in the M-S statistic were calculated as: $R_j^* = \sum_{i=1}^{n} r\left[\sum_{l=1}^{c} r_{ijl}/c\right]$. The computation involves the sum of all the scores for the jth sample, then divided by "c".With duplicates, c= 2.

** All Exact *P* values were lower than 0.0001. Therefore, a comparison was not shown. Exact *P* values were calculated using a Monte Carlo procedure with 10000 iterations. At each "n" value, the data for each protocol and attribute came from the exact same panelists.

Table 5.5 Multiple comparisons test including weighted rank sum differences and total differences across serving protocols and attributes for the different-sample set

		Two Serving Sessions 2SS**				**One Serving Session 1SS**			
				Color					
n	*CV*	X1	X2	X3	Sum	X1	X2	X3	Sum
75	38.0	-123	-135	-258	-516	-137	-135.5	-272.5	-545
70	36.7	-113	-125	-238	-476	-127	-125.5	-252.5	-505
65	35.4	-109	-118	-227	-454	-121	-119.5	-240.5	-481
60	34.0	-102	-108	-210	-420	-111	-109.5	-220.5	-441
55	32.5	-94	-100	-194	-388	-101	-99.5	-200.5	-401
50	31.0	-86	-89	-175	-350	-91	-89.5	-180.5	-361
45	29.4	-77	-80	-157	-314	-81	-79.5	-160.5	-321
40	27.7	-66	-72	-138	-276	-71	-69.5	-140.5	-281
35	25.9	-60	-66	-126	-252	-64	-65.5	-129.5	-259
30	24.0	-50	-56	-106	-212	-56	-54.5	-110.5	-221
25	21.9	-44	-50	-94	-188	-50	-48.5	-98.5	-197
20	19.6	-34	-40	-74	-148	-40	-38.5	-78.5	-157
15	17.0	-26	-32	-58	-116	-30	-28.5	-58.5	-117
10	13.9	-16	-22	-38	-76	-20	-18.5	-38.5	-77
				Sweetness					
n	*CV*	x1	x2	x3	Sum	x1	x2	x3	Sum
75	37.978	-129	-147	-276	-552	-136.5	-144	-280.5	-561
70	36.69	-123	-135	-258	-516	-127.5	-133.5	-261	-522
65	35.356	-113	-125	-238	-476	-118.5	-124.5	-243	-486
60	33.969	-103	-115	-218	-436	-108.5	-114.5	-223	-446
55	32.522	-92	-107	-199	-398	-98.5	-104.5	-203	-406
50	31.009	-82	-97	-179	-358	-88.5	-94.5	-183	-366
45	29.418	-73	-88	-161	-322	-80.5	-83.5	-164	-328
40	27.735	-67	-76	-143	-286	-70.5	-73.5	-144	-288
35	25.944	-59	-65	-124	-248	-67.5	-61.5	-129	-258
30	24.019	-51	-54	-105	-210	-57.5	-51.5	-109	-218
25	21.927	-41	-44	-85	-170	-47.5	-41.5	-89	-178
20	19.612	-31	-34	-65	-130	-37.5	-31.5	-69	-138
15	16.984	-27	-27	-54	-108	-32.5	-25	-57.5	-115
10	13.868	-17	-17	-34	-68	-22.5	-15	-37.5	-75

*The weighted rank sums used in the M-S statistic were calculated as: $R_j^* = \sum_{i=1}^{n} r\left[\sum_{l=1}^{c} r_{ijl}\,/c\right]$. The computation involves the sum of all the scores for the *j*th sample, then divided by "c". With duplicates, c= 2

**All pairs were significantly different (α = 0.05). X1 = R(100%)-R(95%), X2 = R(95%)-R(90%), X3 = R(100%)-R(90%).

5.3.3 Effect of serving protocols for duplicated ranking

The performance of both protocols in the different sample set was comparable for both sweetness and color (measured with M-S statistics in Table 5.4 and rank sum differences in Table 5.5). Although the M-S values were always larger in the 1SS protocol, they did not impact the hypothesis test conclusions given that the lowest M-S value obtained was 33 ($P < 0.0001$) with 10 panelists (sweetness, 2SS). For reference, the Chi-Squared critical value with 2 degrees of freedom for a hypothesis test is 5.991. Thus, diminishing the importance of the small differences found between protocols. This suggest that when differences are very obvious, the serving protocol should not alter the results.

With similar samples, the best protocol (the one showing the higher resolution to differences) depended on the attribute. For sweetness, more sensitivity was achieved with 1SS (higher M-S statistics in 13 out of 14 "n" values). Conversely, the 2SS was more sensitive for color, based on higher M-S statistics for all 14 "n" values (Table 5.2). Exploring the M-S statistics and the sum of the rank sum differences in sweetness showed that the M-S values with 1SS were higher than with 2SS except at n=10, where the M-S statistic of the 1SS protocol was 7.4 (P = 0.0245) with a total difference of -32, whereas the M-S statistic of the 2SS protocol was 11.8 (p = 0.0012) with a total difference of -40. As expected, the largest M-S statistics were found at n=75 and were 35.9 ($P < 0.0001$) with 1SS and 25.2 ($P < 0.0001$) with 2SS, confirming higher sensitivity with increased number of panelists at the same degrees of freedom. Higher test statistics generate lower *P* values either from exact, simulated or chi-squared approximations.

For color, the opposite results were observed; the highest paired differences and M-S statistics were observed using 2SS at every "n". With the degree of difference of samples elicited on panelist perception by the set of 100, 95 and 90 % orange juice, the null hypothesis test

conclusions were not affected by the protocol choice at α= 0.05. Although the null hypothesis was rejected at all panel sizes in both attributes, the differences in M-S statistics depending on the protocols show that when samples are similar, how the duplicates are served can affect the sensitivity. Additionally, for a closer degree of difference, it is possible that the hypothesis tests conclusions are also affected at α= 0.05.

With very different samples, the 1SS had higher M-S values in both attributes; however, the relative impact is negligible since all the null hypothesis tests concluded a rejection with (P < 0.0001). The lowest M-S value observed was 33.0 at n = 10 for sweetness using 2SS. While the largest value was 299.8 also for sweetness, with 75 panelists using 1SS.

5.3.4 Multiple comparisons

Starting with the similar sample set, Table 5.3 shows the weighted (divided by 2 replications) rank sum differences used in multiple comparisons analysis with the M-S experiment-wise error rate test. At each "n" the critical value (α = 0.05) is shared by both attributes and protocols given that in every hypothesis test, "k" and "c" remained constant. Rank sum differences with bold font represent significant paired differences. With three samples, the three possible paired differences between the orange juice samples are represented by X1 = R100%- R95%, X2 = R95%-90%, and X3= R100%-R90%.

In color, all the differences in X1 and X3 were significant regardless of the serving protocol (P < 0.05). In X2, the serving protocol had more influence; with 1SS the rank sum differences were non-significant, and lower than with 2SS. With 2SS, after increasing the panel to n = 55, significant differences were found (X2= -33), also including n= 60 (X2= -36) and 65 (X2= -37); whereas, at n= 70 (X2= -36) and 75 (X2= -36), the differences were almost significant. In contrast, the highest rank sum difference for X2 using the 1SS protocol with similar samples in color was -

13.5 (n= 65. It was expected that X3 showed the largest differences given a 10% difference in orange juice; but it was less obvious to observe that the 1SS protocol would show very low rank sum differences in X2. In general, the 2SS protocol also had lower differences in X1; nevertheless, some hypothesis rejections were achieved. In addition, the differences in X2 and in X1 were more balanced in the 2SS protocol, and not as skewed towards X1 as in 1SS. Although, not necessarily symmetric, similar magnitude of differences was expected between X2 and X1 because both pairs had a 5% difference between samples.

With sweetness, the pattern observed was reversed. The magnitude of the differences was more balanced with the 1SS protocol, whereas the 2SS protocol did not show significant differences for X1. Using 1SS consistently found differences in all pairs starting at n= 15 for X3, n= 20 for X1 and n = 55 for X2. Although, the 2SS protocol found a difference in X2 starting at n= 30, at n= 40 (X2 Diff = 25 < CV = 27.7), the difference was not significant again until the panel was increased to n = 55, which was the lowest number of panelists required to consistently reject the null hypothesis after more panelists were added.

As in the overall null hypothesis tests, the exploration of difference magnitudes in multiple comparisons evidenced that the serving protocol eliciting the largest differences depended on the attribute and the human sense associated with it. In addition, more information was gained since at certain "n" values where both protocols promoted a rejection of the global null hypothesis, the paired comparisons accounting for those differences differed depending on the attribute and serving protocol.

With the different sample set, the 1SS protocol produced higher weighed rank sum differences and M-S statistic values than the 2SS. Nevertheless, the increase in total rank sums at each attribute was 6.5% at most at n= 15 in sweetness (2SS = -108, and 1SS = -115), and could be

as low as 0.83% (2SS = -116, and 1SS = -117), with the same number of panelists in sweetness. Additionally, the values of all rank sums for the 2SS protocol were significant in every test. For example, even at n= 10, the total sum of differences in color was 76 with the 2SS protocol and 77 with the 1SS protocol, with the lowest paired difference being 16 with 2SS, a value higher than the critical value (13.9). The selection of the protocol did not impact the conclusions of the hypothesis tests with different samples as with the similar sample set; although, it could be seen an increase of up to 6% in total differences.

5.4 Discussion

The initial aim of the study was to evaluate if the two serving sessions protocol (2SS) protocol was more adequate for an expectedly more difficult or complex attribute such a sweetness, and the one serving session (1SS) alternative could fit a "simpler" yellow color evaluation. It was shown that color was in general easier to differentiate than sweetness, but it was the color evaluation where the separating the duplicates and allowing a break helped panelists with the detection of differences. Whereas in sweetness, (1SS) helped differentiation. This moves the explanation from attribute complexity to possible specific reasons behind such findings.

The notion that the best serving protocol depended on the attribute and the task complexity can be explained by several reasons that vary between the attributes. Sensitivity to differences in sensory testing using a chemical sense such as taste is affected by the number of physical samples evaluated. Most studies on the effect of the number of samples on sensitivity or statistical power of sensory tests are focused on discrimination testing and not on ranking. In general, when panelists evaluate more samples in discrimination tests of the same cognitive strategy, the sensitivity measured by d' is reduced (Dessirier and O'Mahony, 1998; Rousseau and others, 1998; Rousseau and Others, 1999; Dessirier and others, 1999). The sensitivity reduction can be caused by

adaptation (Ennis and O'Mahony, 1995; O'Mahony, 1986), memory interference (Lau and others 2004) and irritation (Rousseau and O'Mahony, 1999). In discrimination with orange juice models, Cubero and others (1995) found that memory impacts sensitivity more than adaptation, even with paired comparison tests where only two samples per test are tasted. On the other hand, irritation should not impose a difficulty with increased number of samples.

In this study, evaluating both replications in a single session (1SS), thus evaluating six instead of three samples (2SS) showed the opposite effect in sweetness, increasing the resolution of the differences. This was more evident with similar samples than with very different samples. The increase in rank sum differences shows that the 1SS protocol can overcome the previously mentioned adverse factors for this attribute due to a possible cognitive advantage. Posterior interviews with panelists revealed that the closeness and difficulty of some samples (three pairs of twins in a six-sample set), helped separate the samples that actually differed in orange juice concentration. It is then argued that ranking of six samples composed of three pairs of identical samples generated large difficulties to panelists to order the two duplicates of one sample, but created a contrast with the two identical pairs of the other two samples, increasing the ranking resolution. With the data collected it is difficult to quantify the effect of each adverse factor, but it is apparent that harmful effects of increasing the number of samples are less impactful than the cognitive advantage gained by tasting three samples duplicated in the same session. This increase in rank sum differences magnitude could be of similar nature to the increase in correct responses and power gained in the tetrad test over triangle tests when adding an extra sample does not excessively increase the noise in perception (Ennis and Jesionka, 2011; Ennis, 2012; Ishii and others, 2014). Carlisle (2014) reported that panelists valued the forth sample in a tetrad as a confirmatory sample when compared with a triangle test. In this study, panelists reported that they

obtained more insights from the 1SS protocol to mentally group samples before ranking. With three samples and two replications evaluated together (1SS), panelists evaluate two identical aliquots for each of the three samples without knowing it, but gaining insight on what represents a difference in actual percentage of juice

In color, the 2SS protocol exhibited the highest separation between samples. It could be *a priori* hypothesized that a larger number of samples in visual attributes might impart less fatigue or adverse effects possibly showing more power (e.g., the 1SS protocol outperforming 2SS in color ranking); however, it was not the case in this study. The possible causes of the lower sensitivity observed in the 1SS may be linked to chromatic adaptation, a space and time dependent phenomena (Rinner and Gegenfurtner, 2000). In this mechanism, the cones in the retina become less sensitive to a specific wavelength with longer exposure (Fairchild, 2013; Werner, 2014). Ranking six similar samples takes longer time than ranking three samples twice, hence increasing the probability of adaptation. Studies suggested that the adaptation mechanism has fast and slow processes that can start as early as in seconds from exposure and could reach completion within 1 minute (Fairchild and Lennie, 1992; Werner, 2014). On the other hand, memory should not impose a detrimental effect for color ranking since all samples were presented together limiting interruptions to the continuity in perception, (Kinchla and Smyzer, 1967). This study shows that an attribute such as yellow color, in which differences were more easily assessed, panelists can experience negative effects with higher number of samples. On the other hand, sweetness, an attribute where differences were harder to assess, can gain higher resolution to differences with a protocol that has more samples, including both duplicates (1SS). In both evaluations, cognitive and physiological factors influence sensitivity, but the predominant influences in sensitivity of the duplicated ranking appear to be physiological in color and cognitive in sweetness.

5.5 Conclusion

This study showed that the attribute and the complexity of the differences should be considered when selecting a duplicated ranking serving protocol, because different psychological and physiological factors play a role in ranking sensitivity. In this study, duplicated ranking on sweetness gained higher resolution to detect differences when both replications were served jointly in one session (1SS) compared to separately with a break period (2SS). Conversely, color gained higher resolution when each duplicate was presented separately showing that increasing the sample size in color evaluations might not be as simple as conventional wisdom tells. The choice of a protocol for replicated ranking depends not only on degree of difference between samples but also the sense used and stimuli evaluated. Therefore, researchers should test their serving protocols for maximum sensitivity before standardizing practices for continuous testing. It is recommended to test the 1SS protocol for the product and attribute characteristics and opt for the 2SS duplicated ranking only if 1SS does not meet the sensitivity to differences of the 2SS protocol.

5.6 References

Anderson DA. 1988. Some models for overdispersed binomial data. Australian Journal of Statistics 30(2):125-48.

Basker D. 1988. Critical-values of differences among rank sums for multiple comparisons. Food Technology 42(2):79-84.

Bi J. 2009. Computer-intensive methods for sensory data analysis, exemplified by Durbin's rank test. Food Quality and Preference 20(3):195-202.

Bloom DJ, Lee SY. 2016. Sample Dimensionality Effects on d′ and Proportion of Correct Responses in Discrimination Testing. Journal of food science 81(9): S2246–S2251. Brockhoff PB. 2003. The statistical power of replications in difference tests. Food Quality and Preference 14(5): 405-17.

Carabante KM, Alonso-Marenco JR, Chokumnoyporn N, Sriwattana S, Prinyawiwatkul W. 2016. Analysis of Duplicated Multiple-Samples Rank Data Using the Mack–Skillings Test. Journal of food science 81(7):S1791-S1799.

Carlisle SL. 2014. Comparison of Triangle and Tetrad Discrimination Methodology in Applied, Industrial Manner. Master's Thesis, University of Tennessee. Available from: http://trace.tennessee.edu/utk_gradthes/2798. Accessed 2017 March 1.

Christensen ZT, Ogden LV, Dunn ML, Eggett DL. 2006. Multiple comparison procedures for analysis of ranked data. Journal of food science 71(2):S132-S43.

Conover W. 1971. Practical nonparametric statistics. New York, Wiley.

Conover W. 1999. Practical nonparametric statics. John Wiley & Sons, Inc., New York.

Cubero E, Avancini De Almeida TC, O'Mahony M. 1995. Cognitive aspects of difference testing: Memory and interstimulus delay. Journal of Sensory Studies 10(3):307-24.

Dessirier J-M, O'Mahony M. 1998. Comparison of d′ values for the 2-AFC (paired comparison) and 3-AFC discrimination methods: Thurstonian models, sequential sensitivity analysis and power. Food Quality and Preference 10(1):51-8.

Dessirier J, Sieffermann J, O'Mahony M. 1999. Taste discrimination by the 3-afc method: testing sensitivity predictions regarding particular tasting sequences based on the sequential sensitivity analysis model. Journal of sensory studies 14(3):271-87.

Ennis DM, Bi J. 1998. The Beta-Binomial Model: Accounting for inter-trial variation in replicated difference and preference tests. Journal of Sensory Studies 13(4):389-412.

Ennis DM, O'Mahony M. 1995. Probabilistic models for sequential taste effects in triadic choice. Journal of Experimental Psychology: Human Perception and Performance 21(5):1088.

Ennis JM. 2012. Guiding the switch from triangle testing to tetrad testing. Journal of Sensory Studies 27(4):223-31.

Ennis JM, Jesionka V. 2011. The power of sensory discrimination methods revisited. Journal of Sensory Studies 26(5):371-82.

Fairchild MD. 2013. Color appearance models: John Wiley & Sons.

Fairchild MD, Lennie P. 1992. Chromatic adaptation to natural and incandescent illuminants. Vision research 32(11):2077-85.

Friedman M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32(200):675-701.

Hollander M, Wolfe DA, Chicken E. 2013. Nonparametric statistical methods: John Wiley & Sons.

Ishii R, O'Mahony M, Rousseau B. 2014. Triangle and tetrad protocols: Small sensory differences, resampling and consumer relevance. Food Quality and Preference 31:49-55.

Joanes D. 1985. On a rank sum test due to Kramer. Journal of food science 50(5):1442-4.

Kinchla R, Smyzer F. 1967. A diffusion model of perceptual memory. Perception & psychophysics 2(6):219-29.

Lau S, O'Mahony M, Rousseau B. 2004. Are three-sample tasks less sensitive than two-sample tasks? Memory effects in the testing of taste discrimination. Perception & psychophysics 66(3):464-74.

Lawless HT, Heymann H. 2010. Sensory evaluation of food: principles and practices: Springer Science & Business Media.

Mack GA, Skillings JH. 1980. A Friedman-type rank test for main effects in a two-factor ANOVA. Journal of the American Statistical Association 75(372):947-51.

Mennella JA, Colquhoun TA, Bobowski NK, Olmstead JW, Bartoshuk L, Clark D. 2017. Farm to Sensory Lab: Taste of Blueberry Fruit by Children and Adults. Journal of Food Science 82(7): 1713-1719

Meilgaard MC, Carr BT, Civille GV. 2006. Sensory evaluation techniques: CRC press.

O'Mahony M. 1986. Sensory adaptation. Journal of Sensory Studies 1(3-4):237-58.

Rinner O, Gegenfurtner KR. 2000. Time course of chromatic adaptation for color appearance and discrimination. Vision Research 40(14):1813-26.

Rousseau B, Meyer A, O'Mahony M. 1998. Power and sensitivity of the same-different test: comparison with triangle and duo-trio Methods. Journal of Sensory Studies 13(2):149-73.

Rousseau B, Rogeaux M, O'Mahony M. 1999. Mustard discrimination by same–different and triangle tests: aspects of irritation, memory and τ criteria. Food Quality and Preference 10(3):173-84.

Rousseau B, Stroh S, O'Mahony M. 2002. Investigating more powerful discrimination tests with consumers: Effects of memory and response bias. Food Quality and Preference 13(1):39-45.

Stone H, Bleibaum R, Thomas HA. 2012. Sensory evaluation practices: Academic press.

Urbanus BL, Schmidt SJ, Lee SY. 2014. Sensory differences between product matrices made with beet and cane sugar sources. Journal of food science 79(11): s2354-s2361.

Werner A. 2014. Spatial and temporal aspects of chromatic adaptation and their functional significance for colour constancy. Vision research 104:80-9.

Xia Y, Zhong F, O'Mahony M. 2016. Applying Disruptive Preference Test Protocols to Increase the Number of "No Preference" Responses in the Placebo Pair, Using Chinese Consumers. Journal of food science 81(9): S2233–S2239

CHAPTER 6. SUMMARY AND CONCLUSIONS

Duplicated multiple samples ranking tests are not common in the sensory evaluation discipline in part due to the lack of knowledge of appropriate statistical analysis. The main problem with using the traditional Friedman test for analysis of ranking duplicates is a violation of the requirement of independence between blocks. Violations occur when one panelist performs multiple rankings and each ranking test is considered a separate block. Therefore, the initial step in laying the foundation of a duplicated ranking methodology was the evaluation of appropriate statistical tests. In the first study titled: analysis of duplicated multiple-samples rank data using the Mack–Skillings test (M-S, chapter 3), several options were compared for analysis of duplicated preference ranking data, including several alternative analyses with the Friedman test. For example, evaluating replications individually or with the median of both duplicates. The analyses were performed on data obtained from 125 panelists who ranked orange juice model sets with different or similar samples. From that study it was concluded that The Mack-Skillings test can be used in duplicated preference ranking test analysis to increase power and reduce the number of panelists required. Also, whenever possible, if the number of replications is lower than 4, the exact computation or a Monte Carlo simulation approach should be used to estimate *P* values over the chi-squared approximation.

In the second study, a new approach of serving samples of duplicated samples was evaluated for intensity ranking of yellow color of orange juices. The sensitivity to differences elicited on panelists by serving duplicates jointly in one serving session (1SS) was compared with serving them in two serving sessions with a break (2SS). Panelists were less sensitive to differences among very similar samples with the joint session, showing that the increased number of samples produced negative effects.

The third study, evaluated both protocols in intensity ranking of sweetness using the same orange juice models and the same panelists at a different time. For sweetness, serving samples jointly, increased differentiation among very similar samples, showing an opposite conclusion to the one for yellow color ranking, and evidencing a possible cognitive advantage overcoming the possible fatigue, adaptation, or memory effects of a larger sample set. For different attributes, the best protocol was different suggesting that the degree of difference between samples and the attribute characteristics influenced the serving protocol which evoked more sensitivity to differences in intensity ranking. Thus, preliminary studies should determine which protocol suits the attributes and samples of interest for continuous testing.

Although this research showed potential benefits of specific sample serving protocols for yellow color intensity and sweetness, the number of samples of $k = 3$ was not large. More research is needed to understand the effects of larger number of samples (n) in replicated ranking testing. Orange juice samples do not cause irritation; the effects of sensory irritation on the best protocols for sample serving are worthwhile researching. Additionally, other attributes and products should be investigated.

APPENDIX A. ANALYSIS OF DUPLICATED MULTIPLE-SAMPLES RANK DATA USING THE MACK–SKILLINGS TEST IN CHAPTER 3

a. Computer ballot example of preference ranking performed by one panelist on one individual duplicate

Welcome to LSU's Sensory Evaluation Lab

Press the 'Continue' button below to begin the test.

Research Consent Form

I agree to participate in the research entitled “Sensory characteristics of low sodium roasted peanuts containing sodium chloride (NaCl), potassium chloride (KCl) and glycine (Gly)” which is being conducted by Witoon Prinyawiwatkul of the School of Nutrition and Food Science at Louisiana State University Agricultural Center, (225) 578-5188.

I understand that participation is entirely voluntary and whether or not I participate will not affect how I am treated on my job. I can withdraw my consent at any time without penalty or loss of benefits to which I am otherwise entitled and have the results of the participation returned to me, removed from the experimental records, or destroyed. Two hundred consumers will participate in this research. For this particular research, about 5-10 minute participation will be required for each consumer.

The following points have been explained to me:
1. In any case, it is my responsibility to report prior participation to the investigator any food allergies I may have.

2. The reason for the research is to evaluate how consumer liking of low sodium roasted peanuts varies with different concentrations of NaCl, KCl, and Gly. The benefit that I may expect from it is a satisfaction that I have contributed to solution and evaluation of problems related to such examination.

3. The procedures are as follows: three coded samples will be placed in front of me, and I will evaluate them by normal standard methods and indicate my evaluation on score sheets. All procedures are standard methods as published by the American Society for Testing and Materials and the Sensory Evaluation Division of the Institute of Food Technologists.

4. Participation entails minimal risk: The only risk may be an allergic reaction **orange juice, and unsalted crackers. However, because it is known to me beforehand that all those foods and ingredients are to be tested, the situation can normally be avoided.**

5. The results of this study will not be released in any individual identifiable form without my prior consent unless required by law.

6. The investigator will answer any further questions about the research, either now or during the course of the project.

The study has been discussed with me, and all of my questions have been answered. I understand that additional questions regarding the study should be directed to the investigator listed above. In addition, I understand the research at Louisiana State University AgCenter that involves human participation is carried out under the oversight of the Institutional Review Board. Questions or problems regarding these activities should be addressed to Dr. Michael Keenan of LSU AgCenter at 578-1708. I agree with the terms above.

Question # 1.

Your Name:_____

You will be performing two ranking tests.

Before assigning rank (1, 2 or 3) values to the samples please try all three the samples and use crackers and water to cleanse your palate in between samples.

Question # 2.

Please evaluate all samples and **rank** them according to your personal preference.
1st click the sample of your highest preference; **2nd**, click the sample of your intermediate preference and **3rd**, click the sample of your lowest preference.

Rank	**Sample #**
_____	<<Sample1>>
_____	<<Sample2>>
_____	<<Sample3>>

THANK YOU!

b. Counter balanced presentation design of orange juice samples for an individual duplicate

Project: SET 2 (Rep 2 of the similar sample set) Design

Plan:

Description:All Possible Combinations
Description:All Possible Combinations
Type:Quantitative Descriptive
Samples:3
Presented:3
Blocks:1 [Base Block]
X125 [Factor]
=125 [Entire Block]

Options:

Blinding Codes:Constant
Blinding Codes:Constant
Sample Randomization:Yes
Block Randomization:No

Registration:Panelists Will NOT Register
Sample Set DistributionAssign Sample Sets to Panelist on Demand

Sessions:

Number of Sessions:1

Samples:

Sample Number	Product Code	Product Name
1	T100.	Tropicana. 100%
2	T95	Tropicana 95%
3	T90	Tropicana 90%

Blinding Codes for Session 1

Sample Number	Blinding Code	Product Code	Product Name
1	534	T100.	Tropicana. 100%

2	926	T95	Tropicana 95%
3	332	T90	Tropicana 90%

Layout for Session 1 (Example with n= 10 from n= 125)

Sample Set	1	2	3
Sample Set	1	2	3
1	2-926	1-534	3-332
2	1-534	3-332	2-926
3	3-332	2-926	1-534
4	3-332	1-534	2-926
5	3-332	2-926	1-534
6	1-534	3-332	2-926
7	3-332	1-534	2-926
8	3-332	1-534	2-926
9	1-534	3-332	2-926
10	3-332	1-534	2-926

APPENDIX B. SERVING PROTOCOLS FOR DUPLICATED SENSORY RANKING TESTS: SINGLE VERSUS DOUBLE SERVING SESSIONS IN CHAPTER 5

a. Computer ballot example of yellow color intensity ranking by one panelist on one joint duplicate using one serving session

Note: A similar ballot was used to measure sweetness intensity.

Set 6

Question # 1.

Your Name: _____

Please observe the yellow color of all the juice samples, then click continue.

Question # 2.

First. Click the Juice sample (number) with the most intense yellow color.
Then. Continue clicking the second most intense juice sample, then the third, etc...
Finally. The least intense sample will be automatically selected.

Rank	Sample # (Random codes automatically assigned)
_____	<<Sample1>>
_____	<<Sample2>>
_____	<<Sample3>>
_____	<<Sample4>>
_____	<<Sample5>>
_____	<<Sample6>>

THANK YOU!

Do NOT analyze this set again even if you see it appear on your screen.

b. Counter balanced presentation design of orange juice samples for a joint duplicate

Project: SET6 Design

Plan:

Description:	All Possible Combinations
Description:	All Possible Combinations
Type:	Quantitative Descriptive
Samples:	6
Presented:	6
Blocks:	1 [Base Block]
X	75 [Factor]
=	75 [Entire Block]

Options:

Blinding Codes:	Constant
Blinding Codes:	Constant
Sample Randomization:	Yes
Block Randomization:	No
Registration:	Panelists Will NOT Register
Sample Set Distribution	Assign Sample Sets to Panelist on Demand

Sessions:

Number of Sessions:	1

Samples:

Sample Number	Product Code	Product Name
1	T100	Tropicana. 100%
2	T70	Tropicana 70%
3	T40	Tropicana 40%
4	T100.	Tropicana. 100%
5	T70.	Tropicana 70%
6	T40.	Tropicana 40%

Blinding Codes for Session 1

Sample Number	Blinding Code	Product Code	Product Name
1	661	T100	Tropicana. 100%
2	291	T70	Tropicana 70%
3	365	T40	Tropicana 40%
4	175	T100.	Tropicana. 100%
5	677	T70.	Tropicana 70%
6	706	T40.	Tropicana 40%

Layout for Session 1 (Example with n= 8 out of n = 75)

Sample Set	1	2	3	4	5	6
1	3-365	6-706	5-677	4-175	1-661	2-291
2	5-677	1-661	4-175	3-365	6-706	2-291
3	6-706	4-175	5-677	1-661	3-365	2-291
4	6-706	1-661	3-365	2-291	4-175	5-677
5	6-706	3-365	5-677	2-291	1-661	4-175
6	3-365	4-175	6-706	2-291	1-661	5-677
7	6-706	1-661	4-175	5-677	3-365	2-291
8	3-365	1-661	4-175	2-291	5-677	6-706

APPENDIX C. PERMISSION TO PUBLISH CHAPTER 3 WHICH FIRST APPEARED IN THE JOURNAL OF FOOD SCIENCE

This Agreement between Louisiana State Univ -- Witoon Prinyawiwatkul ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

License Number	4216680421199
License date	Oct 26, 2017
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Journal of Food Science
Licensed Content Title	Analysis of Duplicated Multiple-Samples Rank Data Using the Mack–Skillings Test
Licensed Content Author	Kennet Mariano Carabante, Jose Ramon Alonso-Marenco, Napapan Chokumnoyporn, Sujinda Sriwattana, Witoon Prinyawiwatkul
Licensed Content Date	May 30, 2016
Licensed Content Pages	1
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Electronic
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	STATISTICAL AND TECHNICAL METHODOLOGIES FOR DUPLICATED MULTIPLE-SAMPLES PREFERENCE AND ATTRIBUTE INTENSITY SENSORY RANKING TEST
Expected completion date	Dec 2017
Expected size (number of pages)	135

Requestor Location	Louisiana State Univ 201B AFS building School of Nutrition and Food Sci Louisiana State University BATON ROUGE, LA 70803 United States Attn: Witoon Prinyawiwatkul
Publisher Tax ID	EU826007151
Billing Type	Invoice
Billing address	Louisiana State Univ 201B AFS building School of Nutrition and Food Sci Louisiana State University BATON ROUGE, LA 70803 United States Attn: Witoon Prinyawiwatkul

APPENDIX D. COPYRIGHT TRANSFER AGREEMENT FOR CHAPTER 3

Journal of Food Science

Published by Wiley on behalf of Institute of Food Technologists (the "Owner")

COPYRIGHT TRANSFER AGREEMENT

Date: 2016-05-03

Contributor name: Witoon Prinyawiwatkul

Contributor address: School of Nutrition and Food Sciences, Louisiana State University Agricultural

Center, Baton Rouge, LA 70803-4200, USA

Manuscript number: JFDS-2016-0407.R1

Re: Manuscript entitled Analysis of duplicated multiple-samples rank data using the Mack-Skillings test (the "Contribution") for publication in Journal of Food Science (the "Journal") published by Wiley Periodicals, Inc. ("Wiley") Dear Contributor(s):
Thank you for submitting your Contribution for publication. In order to expedite the editing and publishing process and enable the Owner to disseminate your Contribution to the fullest extent, we need to have this Copyright Transfer Agreement executed. If the Contribution is not accepted for publication, or if the Contribution is subsequently rejected, this Agreement shall be null and void.
Publication cannot proceed without a signed copy of this Agreement.

A. COPYRIGHT

1. The Contributor assigns to the Owner, during the full term of copyright and any extensions or renewals, all copyright in and to the Contribution, and all rights therein, including but not limited to the right to publish, republish, transmit, sell, distribute and otherwise use the Contribution in whole or in part in electronic and print editions of the Journal and in derivative works throughout the world, in all languages and in all media of expression now known or later developed, and to license or permit others to do so. For the avoidance of doubt, "Contribution" is defined to only include the article submitted by the Contributor for publication in the Journal and does not extend to any supporting information

submitted with or referred to in the Contribution ("Supporting Information"). To the extent supporting information submitted with or referred to in the Contribution ("Supporting Information"). To the extent that any Supporting Information is submitted to the Journal for online hosting, the Owner is granted a perpetual, non-exclusive license to host and disseminate this Supporting Information for this purpose.

2. Reproduction, posting, transmission or other distribution or use of the final Contribution in whole or in part in any medium by the Contributor as permitted by this Agreement requires a citation to the Journal suitable in form and content as follows: (Title of Article, Contributor, Journal Title and Volume/Issue, Copyright © [year], copyright owner as specified in the Journal, Publisher). Links to the final article on the publisher website are encouraged where appropriate.

B. RETAINED RIGHTS

Notwithstanding the above, the Contributor or, if applicable, the Contributor's employer, retains all proprietary rights other than copyright, such as patent rights, in any process, procedure or article of manufacture described in the Contribution.

C. PERMITTED USES BY CONTRIBUTOR

1. **Submitted Version.** The Owner licenses back the following rights to the Contributor in the version of the Contribution as originally submitted for publication (the "Submitted Version"):

    a. The right to self-archive the Submitted Version on the Contributor's personal website, place in a not for profit subject-based preprint server or repository or in a Scholarly Collaboration Network (SCN) which has signed up to the STM article sharing principles [http://www.stm-assoc.org/stm-consultations/scn-consultation-2015/] ("Compliant SCNs"), or in the Contributor's company/ institutional repository or archive. This right extends to both intranets and the Internet. The Contributor may replace the Submitted Version with the Accepted Version, after any relevant embargo period as set out in paragraph C.2(a) below has elapsed. The Contributor may wish to add a note about acceptance by the Journal and upon publication it is recommended that Contributors add a Digital Object Identifier (DOI) link back to the Final Published Version.

    b. The right to transmit, print and share copies of the Submitted Version with colleagues, including via Compliant SCNs, provided that there is no systematic distribution of the Submitted Version, e.g. posting on a list serve, network (including SCNs which have not signed up to the STM sharing principles) or automated delivery.

2. **Accepted Version.** The Owner licenses back the following rights to the Contributor in the version of the Contribution that has been peer-reviewed and accepted for publication, but not final (the "Accepted Version"):

    a. The right to self-archive the Accepted Version on the Contributor's personal website, in the Contributor's company/institutional repository or archive, in Compliant SCNs, and in not for profit subject-based repositories such as PubMed Central, subject to an embargo period of 12 months for scientific, technical and medical (STM) journals and 24 months for social science and humanities (SSH) journals following publication of the Final Published Version. There are separate arrangements with certain funding agencies governing reuse of the Accepted Version as set forth at the following website: http://www.wiley.com/go/funderstatement. The

Contributor may not update the Accepted Version or replace it with the Final Published Version. The Accepted Version posted must contain a legend as follows: This is the accepted version of the following article: FULL CITE, which has been published in final form at [Link to final article]. This article may be used for non-commercial purposes in accordance with the Wiley Self-Archiving Policy [http://olabout.wiley.com/WileyCDA/Section/id-820227.html].

b. The right to transmit, print and share copies of the Accepted Version with colleagues, including via Compliant SCNs (in private research groups only before the embargo and publicly after), provided that there is no systematic distribution of the Accepted Version, e.g. posting on a list serve, network (including SCNs which have not signed up to the STM sharing principles) or automated delivery.

3. **Final Published Version.** The Owner hereby licenses back to the Contributor the following rights with respect to the final published version of the Contribution (the "Final Published Version"):

a. Copies for colleagues. The personal right of the Contributor only to send or transmit individual copies of the Final Published Version in any format to colleagues upon their specific request, and to share copies in private sharing groups in Compliant SCNs, provided no fee is charged, and further provided that there is no systematic external or public distribution of the Final Published Version, e.g. posting on a list serve, network or automated delivery.

b. Re-use in other publications. The right to re-use the Final Published Version or parts thereof for any publication authored or edited by the Contributor (excluding journal articles) where such re-used material constitutes less than half of the total material in such publication. In such case, any modifications must be accurately noted.

c. Teaching duties. The right to include the Final Published Version in teaching or training duties at the Contributor's institution/place of employment including in course packs, e-reserves, presentation at professional conferences, in-house training, or distance learning. The Final Published Version may not be used in seminars outside of normal teaching obligations (e.g. commercial seminars). Electronic posting of the Final Published Version in connection with teaching/training at the Contributor's company/institution is permitted subject to the implementation of reasonable access control mechanisms, such as user name and password. Posting the Final Published Version on the open Internet is not permitted.

d. Oral presentations. The right to make oral presentations based on the Final Published Version.

4. **Article Abstracts, Figures, Tables, Artwork and Selected Text (up to 250 words).**

a. Contributors may re-use unmodified abstracts for any non-commercial purpose. For online uses of the abstracts, the Owner encourages but does not require linking back to the Final Published Version. **b.** Contributors may re-use figures, tables, artwork, and selected text up to 250 words from their Contributions, provided the following conditions are met:

(i) Full and accurate credit must be given to the Final Published Version.

(ii) Modifications to the figures and tables must be noted. Otherwise, no changes may be made.

(iii) The re-use may not be made for direct commercial purposes, or for financial consideration to the Contributor.

(iv) Nothing herein will permit dual publication in violation of journal ethical practices.

D. CONTRIBUTIONS OWNED BY EMPLOYER

1. If the Contribution was written by the Contributor in the course of the Contributor's employment (as a "work-made-for-hire" in the course of employment), the Contribution is owned by the company/institution which must execute this Agreement (in addition to the Contributor's signature). In such case, the company/institution hereby assigns to the Owner, during the full term of copyright, all copyright in and to the Contribution for the full hereby assigns to the Owner, during the full term of copyright, all copyright in and to the Contribution for the full term of copyright throughout the world as specified in paragraph A above.

For company/institution-owned work, signatures cannot be collected electronically and so instead please print off this Agreement, ask the appropriate person in your
Company/institution to sign the Agreement as well as yourself in the space provided below, and email a scanned copy of the signed Agreement to the Journal production editor. For production editor contact details, please visit the Journal's online author guidelines.

2. In addition to the rights specified as retained in paragraph B above and the rights granted back to the Contributor pursuant to paragraph C above, the Owner hereby grants back, without charge, to such company/institution, its subsidiaries and divisions, the right to make copies of and distribute the Final Published Version internally in print format or electronically on the Company's internal network. Copies so used may not be resold or distributed externally. However, the company/institution may include information and text from the Final Published Version as part of an information package included with software or other products offered for sale or license or included in patent applications. Posting of the Final Published Version by the company/institution on a public access website may only be done with written permission, and payment of any applicable fee(s). Also, upon payment of the applicable reprint fee, the company/institution may distribute print copies of the Final Published Version externally.

E. GOVERNMENT CONTRACTS

In the case of a Contribution prepared under U.S. Government contract or grant, the U.S. Government may reproduce, without charge, all or portions of the Contribution and may authorize others to do so, for official U.S.
Government purposes only, if the U.S. Government contract or grant so requires. (U.S. Government, U.K. Government, and other government employees: see notes at end.)

F. COPYRIGHT NOTICE

The Contributor and the company/institution agree that any and all copies of the Final Published Version or any part thereof distributed or posted by them in print or electronic format as permitted herein will include the notice of copyright as stipulated in the Journal and a full citation to the Journal.

G. CONTRIBUTOR'S REPRESENTATIONS

The Contributor represents that the Contribution is the Contributor's original work, all individuals identified as
Contributors actually contributed to the Contribution, and all individuals who contributed are included. If the Contribution was prepared jointly, the Contributor has informed the co-Contributors of the terms of this Agreement and has obtained their written permission to execute this Agreement on their behalf. The Contribution is submitted only to this Journal and has not been published before, has not been included

in another manuscript, and is not currently under consideration or accepted for publication elsewhere. If excerpts from copyrighted works owned by third parties are included, the Contributor shall obtain written permission from the copyright owners for all uses as set forth in the standard permissions form or the Journal's Author Guidelines, and show credit to the sources in the Contribution. The Contributor also warrants that the Contribution and any submitted Supporting Information contains no libelous or unlawful statements, does not infringe upon the rights (including without limitation the copyright, patent or trademark rights) or the privacy of others, or contain material or instructions that might cause harm or injury. The Contributor further warrants that there are no conflicts of interest relating to the Contribution, except as disclosed. Accordingly, the Contributor represents that the following information shall be clearly identified on the title page of the Contribution: (1) all financial and material support for the research and work; (2) any financial interests the Contributor or any co-Contributors may have in companies or other entities that have an interest in the information in the Contribution or any submitted Supporting Information (e.g., grants, advisory interest in the information in the Contribution or any submitted Supporting Information (e.g., grants, advisory boards, employment, consultancies, contracts, honoraria, royalties, expert testimony, partnerships, or stock ownership); and (3) indication of no such financial interests if appropriate.

H. USE OF INFORMATION

The Contributor acknowledges that, during the term of this Agreement and thereafter, the Owner (and Wiley where Wiley is not the Owner) may process the Contributor's personal data, including storing or transferring data outside of the country of the Contributor's residence, in order to process transactions related to this Agreement and to communicate with the Contributor. By entering into this Agreement, the Contributor agrees to the processing of the Contributor's personal data (and, where applicable, confirms that the Contributor has obtained the permission from all other contributors to process their personal data). Wiley shall comply with all applicable laws, statutes and regulations relating to data protection and privacy and shall process such personal data in accordance with Wiley's Privacy Policy located at: www.wiley.com/go/privacy.

[X] I agree to the COPYRIGHT TRANSFER AGREEMENT as shown above, consent to execution and delivery of the Copyright Transfer Agreement electronically and agree that an electronic signature shall be given the same legal force as a handwritten signature, and have obtained written permission from all other contributors to execute this Agreement on their behalf.

Contributor's signature (type name here): Witoon Prinyawiwatkul

Date: May 3, 2016

SELECT FROM OPTIONS BELOW:

[X] Contributor-owned work [ ] U.S. Government work

Note to U.S. Government Employees

A contribution prepared by a U.S. federal government employee as part of the employee's official duties, or which is an official U.S. Government publication, is called a "U.S. Government work", and is in the public domain in the United States. In such case, Paragraph A.1 will not apply but the Contributor must type his/her name (in the Contributor's signature line) above. Contributor acknowledges that the Contribution will be published in the United States and other countries. If the Contribution was not prepared as part of the employee's duties or is not an official U.S. Government publication, it is not a U.S. Government work.

[ ] U.K. Government work (Crown Copyright) For Crown Copyright this form cannot be completed electronically and should be printed off, signed in the Contributor's signatures section above by the appropriately authorized individual and returned to the Journal production editor by email. For production editor contact details please visit the Journal's

Note to U.K. Government Employees

The Journal production editor by email. For production editor contact details please visit the Journal's online author guidelines. *The rights in a contribution prepared by an employee of a UK government department, agency or other Crown body as part of his/her official duties, or which is an official government publication, belong to the Crown and must be made available under the terms of the Open Government License. Contributors must ensure they comply with departmental regulations and submit the appropriate authorization to publish. If your status as a government employee legally prevents you from signing this Agreement, please contact the Journal production editor.*

[ ] Other

Including Other Government work or Non-Governmental Organization work

Note to Non-U.S., Non-U.K. Government Employees or Non-Governmental Organization Employees **for Other Government or Non-Governmental Organization work this form cannot be completed electronically and should be printed off, signed in the Contributor's signatures section above by the appropriately authorized individual and returned to the Journal production editor by email.** For production editor contact details please visit the Journal's online author guidelines. *If you are employed by the Department of Veterans Affairs in Australia, the World Bank, the World Health Organization, the International Monetary Fund, the European Atomic Energy Community, the Jet Propulsion Laboratory at California Institute of Technology, the Asian Development Bank, or are a Canadian Government civil servant, please download a copy of the license agreement from http://exchanges.wiley.com/authors/copyright-and-permissions_333.html and return it to the Journal Production Editor. If your status as a government or non-governmental organization employee legally prevents you from signing this Agreement, please contact the Journal production editor.*

Name of Government/Non-Governmental Organization:

_____

[ ] Company/institution owned work (made for hire in the course of employment)

For "work made for hire" this form cannot be completed electronically and should be printed off, signed and returned to the Journal production editor by email. For production editor contact details please visit the Journal's online author guidelines. *If you are an employee of Amgen, please download a copy of the company addendum from http://exchanges.wiley.com/authors/copyright-and-permissions_333.html and return your signed license agreement along with the addendum.*

APPENDIX E. LSU AGCENTER INSTITUTIONAL REVIEW BOARD (IRB) EXEMPTION FROM INSTITUTIONAL OVERSIGHT

LSU AgCenter Institutional Review Board (IRB)
Dr. Michael J. Keenan, Chair
School of Human Ecology
209 Knapp Hall
225-578-1708
mkeenan@agctr.lsu.edu

Application for Exemption from Institutional Oversight

All research projects using living humans as subjects, or samples or data obtained from humans must be approved or exempted in advance by the LSU AgCenter IRB. This form helps the principal investigator determine if a project may be exempted, and is used to request an exemption.

- Applicant, please fill out the application in its entirety and include the completed application as well as parts A-E, listed below, when submitting to the LSU AgCenter IRB. Once the application is completed, please submit the original and one copy to the chair, Dr. Michael J. Keenan, in 209 Knapp Hall.

- A Complete Application Includes All of the Following:
  - **(A)** The original and a copy of this completed form and a copy of parts B through E.
  - **(B)** A brief project description (adequate to evaluate risks to subjects and to explain your responses to Parts 1 & 2)
  - **(C)** Copies of all instruments and all recruitment material to be used.
    - If this proposal is part of a grant proposal, include a copy of the proposal.
  - **(D)** The consent form you will use in the study (see part 3 for more information)
  - **(E)** Beginning January 1, 2009: Certificate of Completion of Human Subjects Protection Training for all personnel involved in the project, including students who are involved with testing and handling data, unless already on file with the LSU AgCenter IRB.
    Training link: (http://grants.nih.gov/grants/policy/hs/training.htm)

1) Principal Investigator: Witoon Prinyawiwatkul **Rank:** Professor **Student?** No
School of Nutrition and Food Sciences **Ph:** 8-5188
E-mail: wprinyawiwatkul@agcenter.lsu.edu **and** wprinya@lsu.edu

2) Co-Investigator(s): please include department, rank, phone and e-mail for each NONE

- If student as principal or co-investigator(s), please identify and name supervising professor in this space

3) Project Title: Consumer Acceptance and Perception of New and Healthier Food Products
4) Grant Proposal?(yes or no) NO **If Yes, Proposal Number and funding Agency**
Also, if Yes, either: **this application completely matches the scope of work in the grant Y/N**
OR
more IRB applications will be filed later Y/N
5) Subject pool (e.g. Nutrition Students) LSU Faculty, Staff, Students and off-campus consumers

- Circle any **"vulnerable populations" to be used:** (children<18, the mentally impaired, pregnant women, the aged, other). Projects with incarcerated persons cannot be exempted. **NONE**

6) PI signature ****Date** 3-12-2015 (no per signatures)
****I certify that my responses are accurate and complete.** If the project scope or design is later changed I will resubmit for review. I will obtain written approval from the Authorized Representative of all non-LSU AgCenter institutions in which the study is conducted. I also understand that it is my responsibility to maintain copies of all consent forms at the LSU AgCenter for three years after completion of the study. If I leave the LSU AgCenter before that time the consent forms should be preserved in the Departmental Office.

Committee Action: Exempted Not Exempted ✓ **IRB#** HE15-9
Reviewer Michael Keenan **Signature** Michael Keenan **Date** 3-16-2015

VITA

Kennet Mariano Carabante-Ordoñez was born in August, 1988 in Tegucigalpa, Honduras. He obtained his Bachelor of Science degree in Agro-Industrial engineering from Zamorano University in 2008. In 2010 he worked as a processing researcher at the Honduran Institute of Coffee before moving to Louisiana State University as a visiting scholar. In 2013 he obtained his master's in Food Science from LSU. Afterward, he continued on to the PhD program in Food Science heavily focused on Sensory and Consumer Sciences with a minor in Applied Statistics, which he expects to complete in December 2017.